DATE DUE

For each day's delay after the due date a fine of 3 P. per Vol. shall be charged for the first week, and 25 P. per Vol. per day for subsequent days.

Borrower's No.	Date Due	Borrower's No.	Date Due

GIPNLK—5/PPND/64—27-10-64—5,000.

COLLEGE OUTLINE SERIES

A. W LITTLEFIELD, *General Editor*

ACCOUNTING, Elementary . . . . BAUER DARBY
ALGEBRA, College . . . . . MOORE
ANCIENT HISTORY . . . . . HYMA
ANCIENT, Medieval, Modern History . . . RICKARD-HYMA

BACTERIOLOGY, Principles of . . . . . BRYAN
BIOLOGY, General . . . . . ALEXANDER
BOTANY, General . . . . . FULLER
BUSINESS LAW . . . BABB MARTIN

CALCULUS, The . . . . . OAKLEY
CHEMISTRY, First Year College . . . . . LEWIS
CHEMISTRY, Mathematics for General . . . FREY
CHEMISTRY, Organic . . . . DEGERING, et al
CORPORATION FINANCE . . . . . HAROLD

DOCUMENTED PAPERS, Writing . . . . . HUBBELL

ECONOMICS, Principles of . . . . . JAMES
EDUCATION, History of . . . . . THOMPSON
ENGINEERING DRAWING . . . . LOMBARDO-JOHNSON
ENGLAND, History of . . . . . RICKARD
EUROPE, 1500 1848, History of . . . . . LITTLEFIELD
EUROPE, 1815 1946, History of . . . . . LITTLEFIELD
EXAMINATIONS, How to Write Better . . . . HOOK

FRENCH GRAMMAR . . . . . DU MONT

GEOLOGY, Principles of . . . . . FIELD
GEOMETRY, Plane, Problems in . . . HORBLIT-NIELSEN
GERMAN GRAMMAR . . . . . GREENFIELD
GOVERNMENT, American . . . . . SAYRE
GRAMMAR, English, Principles of . . . . . CURME

HYDRAULICS for Firemen . . . . . THEOBALD

JOURNALISM, Survey of . . . . . MOTT, et al

[*Continued on next page*]

[*Continued from preceding page*]

LATIN AMERICA, History of . . . . . WILGUS-D'EÇA
LATIN AMERICA in Maps . . . . . WILGUS
LATIN AMERICAN Civilization, Readings in . WILGUS
LATIN AMERICAN Economic Development . . WYTHE
LITERATURE, American . . . . . CRAWFORD, et al
LITERATURE, English, Dictionary of . . . . . WATT
LITERATURE, English, To Dryden . . OTIS-NEEDLEMAN
LITERATURE, English, Since Milton . . NEEDLEMAN-OTIS
LOGARITHMIC and Trigonometric Tables . . . NIELSEN

MIDDLE AGES, 300-1500, History of . . . . MOTT-DEE
MUSIC, History of . . . . . MILLER

PHILOSOPHY An Introduction . . . RANDALL-BUCHLER
PHILOSOPHY, Readings in . . . . . RANDALL, et al
PHYSICS, First Year College . . . . . BENNETT
POLITICAL SCIENCE . . . . . JACOBSEN-LIPMAN
POLITICS, Dictionary of American . . SMITH-ZURCHER
PORTUGUESE GRAMMAR . . . . D'EÇA-GREENFIELD
PSYCHOLOGY, Educational . . . . . PINTNER, et al
PSYCHOLOGY, General . . . . . FRYER-HENRY

SHAKESPEAREAN NAMES, Dictionary of . . IRVINE
SHAKESPEARE'S PLAYS, Outlines of . . WATT, et al
SLIDE RULE, Practical Use of . . . . . BISHOP
SOCIOLOGY, Principles of . . . . . LEE, et al
SPANISH GRAMMAR . . . . . GREENFIELD
STATISTICAL METHODS . . . . . ARKIN-COLTON
STUDY, Best Methods of . . . . . SMITH-LITTLEFIELD

TRIGONOMETRY . . . . . NIELSEN-VANLONKHUYZEN
TUDOR and Stuart Plays, Outlines of . . . HOLZKNECHT

UNITED STATES, To 1865, History of . . . . KROUT
UNITED STATES, Since 1865, History of . . . KROUT
UNITED STATES in Second World War . . . . HARRIS

WORLD Since 1914, History of . . . . . LANDMAN

ZOOLOGY, General . . . . . ALEXANDER

AN OUTLINE OF STATISTICAL METHODS

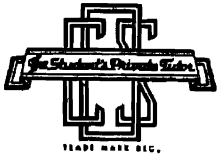
AS APPLIED TO
ECONOMICS, BUSINESS, EDUCATION,
SOCIAL AND PHYSICAL SCIENCES, ETC

By

HERBERT ARKIN
The College of the City of New York
and
RAYMOND R. COLTON
The College of the City of New York

WITH A PREFACE BY
JUSTIN H. MOORE, Professor of Law
School of Business and Civic Administration
The College of the City of New York

Fourth Edition



New York

BARNES & NOBLE, Inc.

FIRST EDITION — 1934
SECOND EDITION — 1935
THIRD EDITION — 1938
FOURTH EDITION — 1939
REPRINTED — 1942
REPRINTED NOVEMBER, 1943
REPRINTED MAY, 1946
REPRINTED, MARCH 1947

Printed in the United States of America

Preface

In Europe it has long been the practise to publish small books known as *manuals*, giving a condensed and succinct treatment of the subject as an aid to the student in summarizing the more elaborate and detailed contents of a large textbook. Halfway between an article in an encyclopedia and the often discursive texts studied in the classroom, these manuals fulfil a real need; they winnow fundamental principles from a mass of material, give with incisive clarity the contours of the ground already covered and leave with the reader a definite framework which later he may fill in by further practical contact with the subject in real life. In America there has been a noticeable gap in educational literature which now happily will be filled in part by the excellent series of small books published by Barnes & Noble, Inc.

The present volume on statistics does not aim to be a comprehensive treatise on the subject. On the contrary it gives the distilled essence of material which might well require one or more large volumes for a full discussion. For that very reason it ought to be a most useful tool in the hands alike of students and people actually engaged in statistical work, wherever the particular field of activity may happen to lie. The formulas and examples given in it will be ample for the needs of most workers, whether they be concerned with financial, industrial, commercial, social, or educational statistics. Of necessity there is a minimum of verbiage; no formula is included that has not practical applications; the mathematical aspects are not stressed; the philosophy of the subject and many recondite byways are not explored. Thus the reader is spared the need of hunting back and forth to find the special help which he needs on his concrete problems in daily life. To all statistical workers this little volume will be as indispensable as an adding machine.

JUSTIN H. MOORE, Professor of Law
School of Business and Civic Administration
The College of the City of New York

Table of Contents

PUBLISHER'S NOTE

In this edition we have included eighteen standard textbooks in the tabulated bibliography table. The following list gives the author, title, and publisher of all the books referred to in the tables on the next two pages.

Camp, B. H., *Mathematical Part of Elementary Statistics*, Heath, 1931.

Chaddock, R. E., *Principles and Methods of Statistics*, Houghton Mifflin, 1925.

Croxton, F. E. and Cowden, D J, *Practical Business Statistics*, Prentice-Hall. 1934.

Crum, W. L, Patton, A C. and Tebbut, A. R., *An Introduction to Economic Statistics*, McGraw-Hill, Rev. 1938.

Davies, G. R. and Crowder, W. F., *Methods of Statistical Analysis in Social Sciences*, Wiley, 1933.

Day, E. E., *Statistical Analysis*, Macmillan, 1930.

Garrett, H. E., *Statistics in Psychology and Education*, Longmans, Rev. 1937.

Harper, F., *Elements of Practical Statistics*, Macmillan, 1930.

Holzinger, K. J, *Statistical Methods for Students in Education*, Ginn, 1928.

Jerome, H, *Statistical Method*, Harper, 1924.

Kelley, T. L, *Statistical Method*, Macmillan, 1923.

Mills, F C., *Statistical Methods*, Holt, 2nd ed. 1938.

Odell, C. W., *Statistical Method in Education*, pp. 14-27, Appleton-Century, 1935.

Richardson, C. H, *An Introduction to Statistical Analysis*, Harcourt, 1934.

Riggleman, J. R. and Frisbee, I. N., *Business Statistics*, McGraw-Hill, Rev. 1938.

Rugg, H. O., *Statistical Methods Applied to Education*, Houghton Mifflin, 1917.

Thurstone, L. L., *Fundamentals of Statistics*, Macmillan, 1925.

Yule, G. U., *An Introduction to the Theory of Statistics*, Griffen (London) Rev. 1937.

Quick Reference Tabl

All figures refe

CHAPTER IN THIS OUTLINE	TOPIC	LAMP	CHADDOCK	CROXTON and COWDEN	CRUM et al	DAVIES	DAY	GARRET
I	Statistical Series		44 80	151 153	3 11	4 33	36-47 118 133	4 9
II	Frequency Distribution—Central Tendency—Arithmetic Mean		81 105	153 172	76 181	33 38	134 139	17 19 26 29
III	Frequency Distribution—Central Tendency (continued)	36 42	107 148	176 202	182 192	38 57	140 162	20 26 29
IV	Frequency Distribution—Dispersion and Skewness	43 46	150 173	204 219	193 207	65 85	163 179	33 61
V	Time Series Analysis—Trend	102 104	206 320	262 282	298 300	138 139 153 156	231 257	
VI	Time Series Analysis The Least Squares Method Linear	104 111	320 334	313 325	300 316	133 137	258 263	
VII	Time Series Analysis Non Linear	112 126	335 339	326 338	316 325	139 153 156-178	263 280	
VIII	Time Series Analysis—Seasonal and Cyclical		349 366	286-310	326 362	189 219	281 312	
IX	Correlation—Linear	179 17 [illegible] 290	48 290	405 4[illegible]	230 252 363 370	226 250	80 210 313 327	251 2 289 31
X	Correlation—Non Linear Multiple Partial	[illegible]	[illegible] 30[illegible]	431 495	253 262	250 280	201 206	393 464
XI	Correlation of Attribute	307 314						366 35
XII	The Normal Curve	97 [illegible] 199	7 228	241 257	208 214	292 306		98 177
XIII	Theory of Sampling	230 273	98 246	22 237	[illegible] 5 220	298 300		194 250
XIV	Index Numbers		72 205	362 397	263 297	91 123	328 367	
XV	Further Analysis of the Frequency Distribution	14 30			221 229			
XVI	Collection of Data		371 395	22 37	38 59			
XVII	Statistical Tables		[illegible] 417	39 58	60 07			
XVIII	Graphic Presentation		[illegible] 445	[illegible] 129	91 154		48 113 211 230	
XIX	Special Techniques in Education Psychology and Biology	8 10		234 239				

See preceding page

Standard Textbooks

pages

ARPER	HOLZ INGER	JEROME	KELLEY	MILLS	ODELL	RICHARD-SON	RIGGLE MAN	RUGG	THUR STONE	YULE
364	1 8		1 8	50 85	14 27	15 41	132 163	4 26 74 81		82 109
110	78 85	109 117	44 53	96 109	32-62 66 76	43 53	164 170	81 100 114 126	1 17 67 77	112 119
2 121	85 96	117 133	54 68	109 136	77 114	53 71	170 181	100-114 126-147	78-85	120 133
152	101 139	146 164	70 82	137 160	115 141	77 95	215 228	149 178	86 123	134 153
168	317 325	224 228		225 246			270-297	178 179	18 28	
178	154 161 321 322	228 233		246 253		113 133	297 300	248 249	51-64	
287	324 337			53 283		169 200	300 310		29 37	
		234 255		284 304			313 359			
0 56	1 1 173	263 287	151 182	402	133 708 237 249	145 162	244 264	232 270 294 296	187 223	196 226
145 1 276	7 186 231 315	286	185 310	304 454 531 597	2 0 308	201 203	264	276 306	224 228	241 252 261 287
	256 278				309 324			292 309		
4 156	190 229	62 181	94 106 109 150	425 435	53 64 378 396	207 249		191 227	126-148	169 187
13 6 161	231 45 245 254	171 177	82 94	452 489 528 539	326 376	251 271		227 231 270 274	161 186	332 394
283		83 223	331 347	161 224 305 324			184 214			
	338 344			435 444 448 450		97 111				154 168
	9 28	290 326			14 31		14 47	28 73 310 360		
	31 34 36-45	48 49		62 63			48 72	87 94 310 360		
93		0 108	9 43	8 49	416 434		76 131	310 360	47 49	
	Entire Volume				Entire Volume		360-634	Entire Volume		

nplete list of titles

THE GREEK ALPHABET

A	α	alpha	O	o	omicron
B	β	beta	Π	π	pi
Γ	γ	gamma	P	ρ	rho
Δ	δ	delta	Σ	σ	sigma
E	ϵ	epsilon	T	τ	tau
Z	ζ	zeta	Υ	υ	upsilon
H	η	eta	Φ	ϕ	phi
Θ	θ	theta	X	χ	chi
I	ι	iota	Ψ	ψ	psi
K	κ	kappa	Ω	ω	omega
Λ	λ	lambda			
M	μ	mu			

CHAPTER I

STATISTICAL SERIES

Definition of Statistical Method

Statistical Method is a technique used to obtain, analyze and present numerical data.

Elements of Statistical Technique

The elements of statistical technique include the:

1. Collection and assembling of data.
2. Classification and condensation of data.
3. Presentation of data in:
    a. textular form.
    b. tabular form.
    c. graphic form.
4. Analysis of data.

Characteristics and Limitations of Statistical Methods

1. Statistical method is the only means for handling large masses of numerical data.

2. Statistical technique applies only to data which are reducible to quantitative form.

3. Statistical technique is *objective*. The results, however, cannot but be affected by the necessarily *subjective* interpretation

4. Statistical technique is the same for the social as for the physical sciences; i.e., economics, education, sociology and psychology as contrasted with biology, chemistry and astronomy. Method and technique apply alike to these divergent fields.

Statistical Series

In order to analyze numerical data, it is first necessary to arrange them systematically. The data may be arranged in a number of different ways. Technically an arrangement is called a distribution or series. An example of each type of distribution is shown below:

When data are grouped according to:	*The resulting series is called a:*
1. Magnitude	Frequency distribution
2. Time of occurrence	Time series
3. Geographic location	Spatial distribution

In addition there is a number of special types of distributions in which the data may be arranged by *kind* or by *degree*.

The Frequency Distribution

Definition

The frequency distribution is an arrangement of numerical data according to size or magnitude.

Construction

A frequency distribution is constructed in the following manner:

1. Using the **range** of the data (the interval between the highest and the lowest figure) as a guide, the data are divided into a number of convenient sized groups. The groups are called **class intervals** (compare table 1, column 1).[1]

The size of the class interval is dependent upon the number of values to be included in the distribution. The range of the values (difference between the highest and lowest values) is determined and is divided by the number of class intervals desired. The resulting size is rounded off. Few class intervals are used when a limited number of values are included and a large number when the distribution is to be compiled from many values. The most efficient number of class intervals usually lies between ten and twenty groups.

Other requirements for the determination of the class interval are

a. The class intervals should not overlap; 0 – 4.9, 5 – 9.9 etc., should be used in preference to 0 – 5, 5 – 10, etc.

b. When the values tabulated coincide with the integers or with selected values, these values should generally constitute the midpoints of the groups.

c. When possible the class intervals should be of a uniform size.

2. The groups are then placed in a column with the lowest class interval at the top and the rest of the class intervals following according to size.
3. The data are then scored. Each figure is checked once next to the class interval into which it falls (see tally, table 1).[2]

Graphic Presentation of Frequency Distribution

If two lines are drawn perpendicular to one another and are divided according to a scale of values, given data may be represented by reference to the scale The horizontal line is known as the X **axis** and the vertical line as the Y **axis.** If the values for

[1] As a preliminary step the raw data may be arranged according to size. The series is then called an array.

[2] In scoring data an efficient procedure is to connect the first and fourth score by the fifth. Using this procedure totals are obtained merely by adding the resulting units multiplying by five and adding the odd scores (see Table 1). A tally of this type saves time in counting frequencies and also eliminates possible inaccuracies

Table 1 — Score Sheet
City Tax Rate of "True" Valuation in 261 Cities in the United States, 1927

Rate Per Thousand Dollars (Class Intervals)	Tally	Total Number of Cities (frequency)
4 – 7.99	𝍸	5
8 – 11.99	𝍸 𝍸 𝍸	15
12 – 15.99	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 I	46
16 – 19.99	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 III	68
20 – 23.99	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 III	58
24 – 27.99	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 II	32
28 – 31.99	𝍸 𝍸 𝍸 𝍸 II	22
32 – 35.99	𝍸 𝍸	10
36 – 39.99	II	2
40 – 43.99	II	2
44 – 47.99		0
48 – 51.99	I	1
		261

Source United States Department of Commerce, *Financial Statistics of Cities*, 1927, Table 23.

a point are given the point may then be located on the graph. For example, in figure 1 the point $X = 2$, $Y = 3$ may now be located at the point marked *a*.

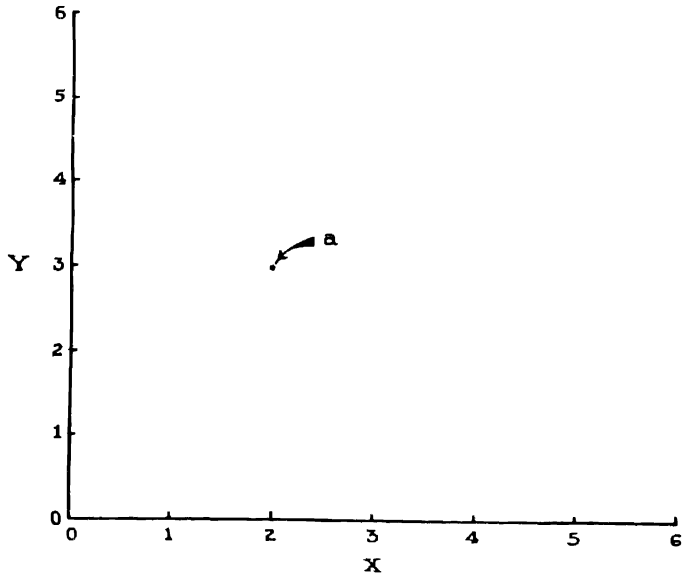


Fig. 1—Location of Plotted point $X = 2$, $Y = 3$.

If the two axes are now marked in the units of the given data the frequency distribution may be presented graphically (pictorially).

1. The class interval grouping which will be termed the **independent variable** is placed on the X (horizontal) axis and the frequency or **dependent variable** is placed on the Y (vertical) axis.[1]
2. The number of cases (frequency) is plotted at the midpoint of each respective class interval at the appropriate horizontal level as indicated by the scale on the Y axis.[2]
3. When connected the plotted points form a **frequency polygon.**
4. Rectangles may be constructed by using as the width the size of the class interval and as the height the frequency in each class interval. The rectangles form an **histogram** (also known as a **rectangular frequency polygon**).

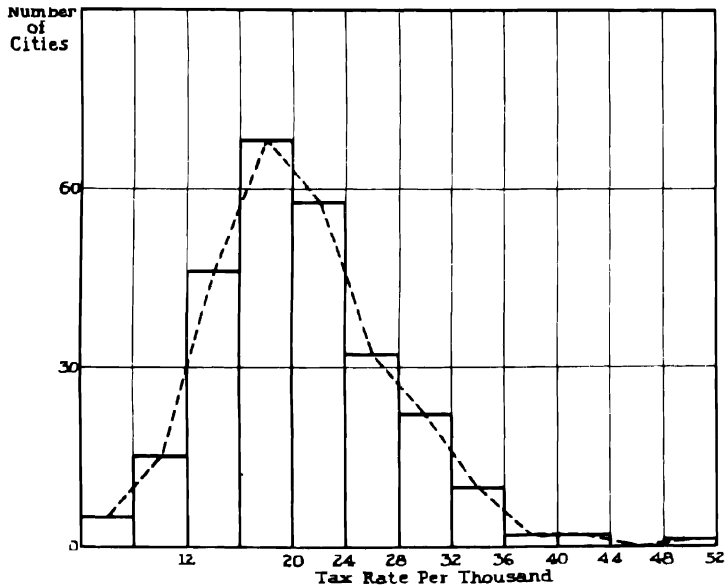


Fig. 2—Distribution of Tax Rates on "True" Valuation for 261 Cities in the United States, 1927.

Cumulative Frequency Distribution

A distribution in which the frequencies are cumulated is known as an **ogive.** Examples of the ogive are shown in tables 2a and 2b.

[1] The distance along the X axis is called the abscissa while that along the Y axis is called the ordinate.

[2] A [illegible] otional background or ruling may be used to aid in this work.

Table 2a — "Less Than" Ogive
Distribution of Wholesale Sales by Size of Firm for the United States, 1930

Size of Firm (Thousands of Dollars of Sales)	Number of Firms
Less than 25	14,235
Less than 50	23,568
Less than 100	36 154
Less than 200	49,667
Less than 300	57,081
Less than 400	61,701
Less than 500	64,716
___ than 1,000	71,453
Less than 25,000	76,600

Source: United States Department of Commerce 1930 Census.

In this table a "less than" ogive is used. This distribution may easily be converted into an "and over" ogive by cumulating the items on an "and over" basis as in table 2b.

Table 2b — Distribution of Farms in New England by Size, 1930

Size in Acres	Number of Farms
0 and over	124,925
20 and over	104,948
50 and over	84,664
100 and over	54,924
175 and over	24,628
260 and over	10,494
500 and over	2,165
1000 and over	392
5000 and over	0

Source. United States Department of Commerce, 1930 Census.

Analysis

Unless a mass of data is grouped it is unwieldly and in many cases impossible to analyze. In a frequency distribution the data are arranged so that with the application of further techniques analysis of the data is made possible. In itself the mere grouping of data does not present an analysis.

Types of Frequency Distributions

The more usual types of distributions are shown below. In addition to these there are a number of unusual types such as the bimodal curve (two peaks), the "j" and inverted "j" curves (curves shaped to resemble a "j"), etc.

1. **Symmetrical distribution:** The "normal" curve is the best known example of a symmetrical distribution.

2. **Skewed distribution:** Most frequency distributions extend further in one direction than in the other. This type is known as a **skewed distribution.** It is, of course, identified by a lack of symmetry.
   a. The **right** (positively) **skewed distribution** is caused by the extremes in the higher values distorting the curve towards the right.

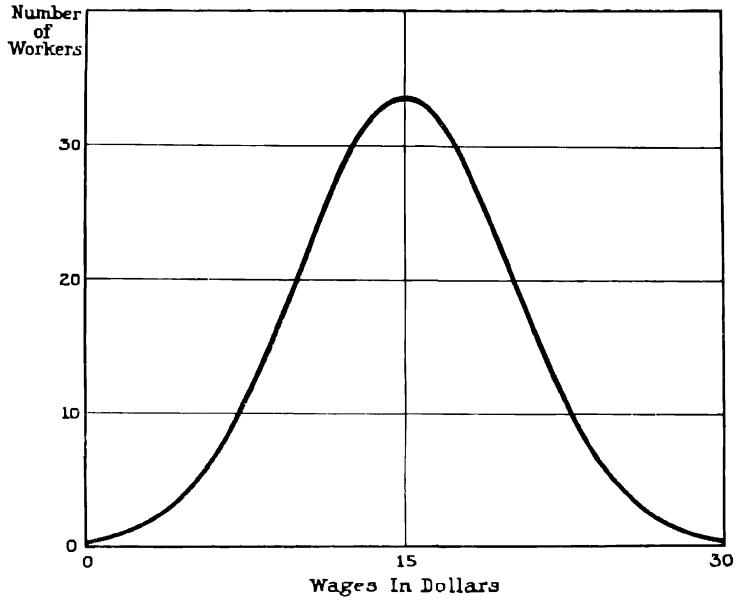


Fig. 3—Hypothetical "Normal" Distribution of Wages in a Factory.

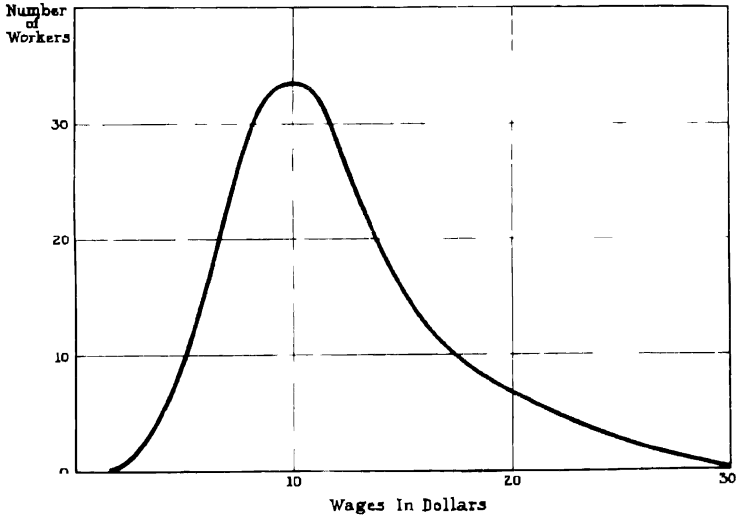


Fig. 4—Hypothetical Right Skewed Distribution of Wages in a Factory.

b. The **left (negatively) skewed distribution,** a less common type, is caused by extremes in the lower values which distort the curve towards the left.

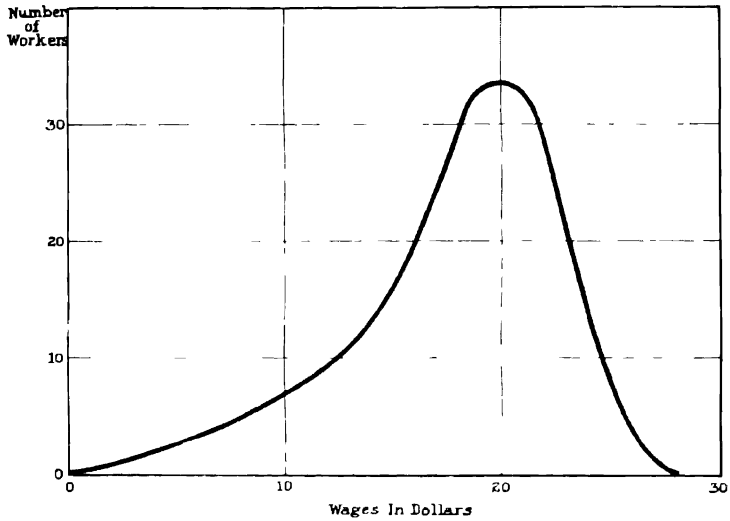


Fig. 5—Hypothetical Left Skewed Distribution of Wages in a Factory.

Characteristics of the Frequency Distribution:[1]

Natural, economic and sociological data show a distinct tendency to group about a given point.

This grouping tendency gives rise to a peak which always occurs in frequency distributions. The location of the peak or *point of central or common tendency* is one of the characteristics which may be measured. In the graph below the two distributions are identical in nature except that the points of central tendency are located at different positions on the scale.

The tendency of a group of values to cluster about a central point makes possible the use of a typical value to describe the mass of data. The location of this point of central tendency may be measured by the average.

[1] Frequency distributions are commonly divided into two types, continuous and discrete (non continuous). In the *continuous* series it is possible to have every size of the variables between given limits, i.e., in the distribution of weights, or of ages of individuals where any size is possible. In the *discrete series* only limited gradations are possible, as in, for instance, the distribution of numbers of pupils in public schools in the United States. Fractions of a unit cannot appear in this series.

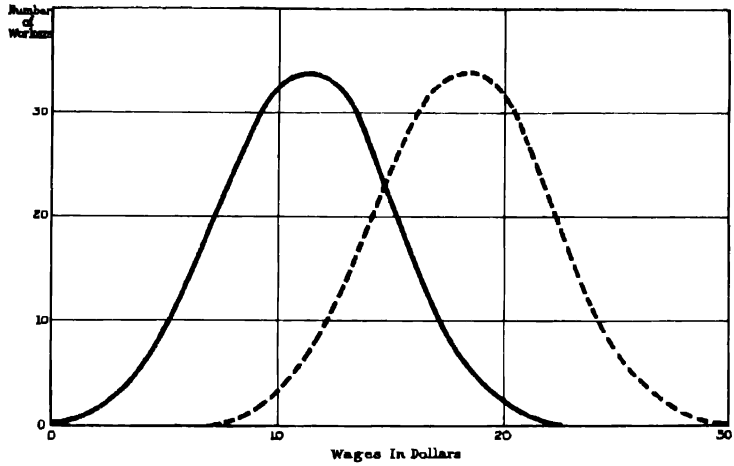


Fig. 6—Hypothetical Distribution of Wages in Two Factories.

Dispersion

In figure 7 the distributions are identical in character. The values of the items included in curve *a*, however, vary to a greater degree than curve *b*. The *degree* of variation differs from curve to curve and is known as **dispersion.** Dispersion, then, may be defined as the variation in size occurring among the various items constituting the series.

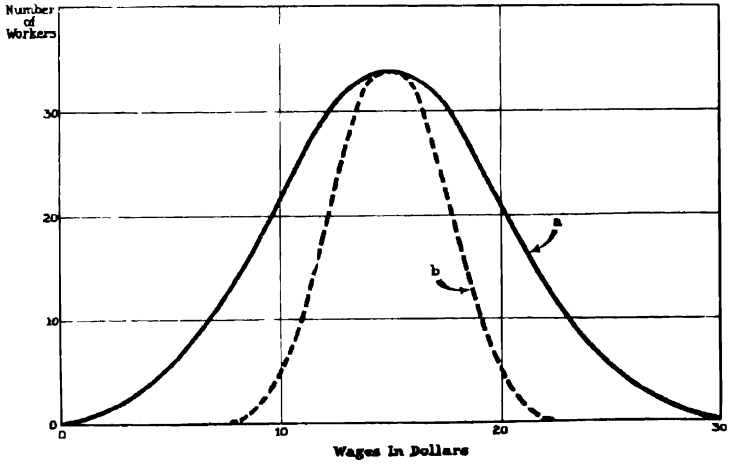


Fig. 7—Hypothetical Distribution of Wages in Two Factories.

Skewness

The distributions in figure 8 differ in that curve *a* is symmetrical while curve *b* is not. The lack of symmetry is known as **skewness.**

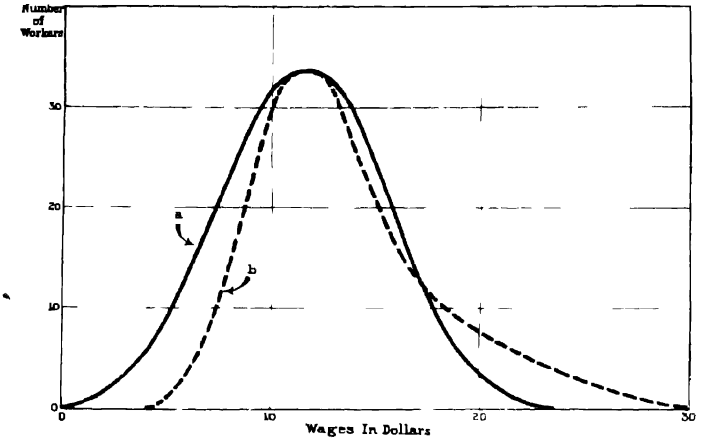
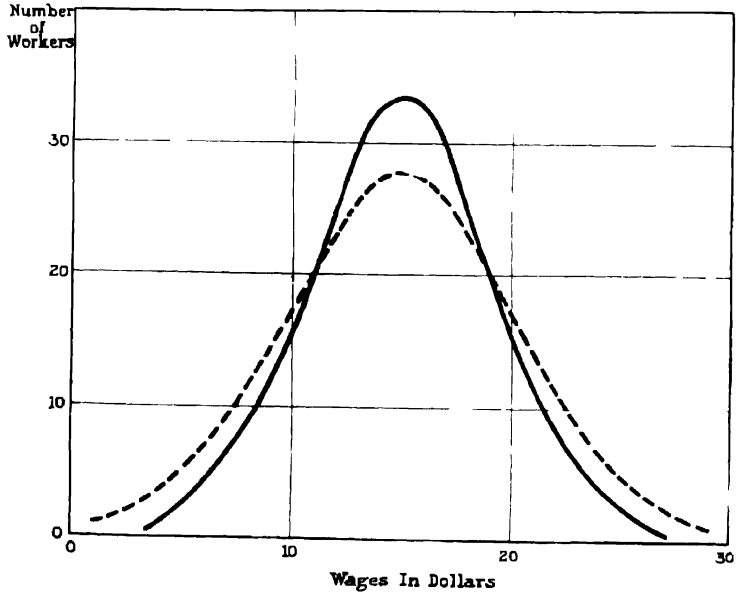


Fig. 8—Hypothetical Distribution of Wages in Two Factories.

Kurtosis

The two curves in figure 9 differ in their degree of "peakedness." This characteristic is known as **kurtosis.**



Source: See table 2

Fig. 9—Hypothetical Distribution of Wages in Two Factories.

ADDITIONAL BIBLIOGRAPHY*

BOWLEY, ARTHUR *w.*, *Elements of Statistics*, pp. 2–13, P. S. King & Son, London, 1907.

DAVIES, GEORGE R., & YODER, DALE, *Business Statistics*, pp. 1–4; 23–29. John Wiley & Sons, New York, 1937.

FISHER, R. A., *Statistical Method for Research Workers*, pp. 1–40. Oliver & Boyd, Edinburgh, 1932.

ODELL, C. W., *Educational Statistics*, Ch. 1. Century Co., New York, 1925.

RIETZ, H. L. (Editor), *Handbook of Mathematical Statistics*, pp. 20–23. Houghton Mifflin Co., New York, 1924.

SUTCLIFFE, WILLIAM G., *Statistics for the Business Man*, pp. 1–16. Harper & Bros., New York, 1930.

ZIZEK, FRANK, *Statistical Averages*, pp. 7–24. Henry Holt & Co., New York, 1930.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER II

ANALYSIS OF THE FREQUENCY DISTRIBUTION CENTRAL TENDENCY-ARITHMETIC MEAN

Measures of Common Tendency-Averages

An average is a typical value which tends to sum up or describe the mass of data. It also serves as a basis for measuring or evaluating extreme or unusual values. The average is a measure of the location of central tendency.

Kinds of Averages

The most important kinds of averages are:

1. The arithmetic mean (Discussed in Chap. II)
2. The median (Discussed in Chap. III)
3. The mode " " "
4. The geometric mean " " "
5. The quadratic mean " " "

The Arithmetic Mean

Due to ease of computation and long usage the **arithmetic mean** is the best known and most commonly used of all the averages.

Methods of Calculating the Arithmetic Mean

Ungrouped Data

The arithmetic mean of a small group of individual items may be obtained by adding all items together and dividing the total by the number of items used.

The computation of the arithmetic mean is expressed in formula form as:[1]

$$\overline{X} = \frac{\Sigma\ (X)}{N}$$

where

$\overline{X}$ = arithmetic mean
Σ = symbol meaning "sum of"[2]
X = data expressed as individual items
N = number of items

[1] This formula is also given in various textbooks, using different symbols, as:

(a) $$M = \frac{1}{N}\Sigma\ (X)$$

(b) $$M = \frac{\Sigma\ (m)}{N}$$

[2] The symbol Σ is a Greek capital letter Sigma

Grouped Data

Where the arithmetic mean of a considerable number of items is to be computed the method outlined above is generally too laborious and too subject to error. With increasing numbers of items the simple problem of addition can become difficult to the point of physical impossibility. For instance, if the arithmetic mean is to be applied to data containing forty or fifty thousand items, correct addition of the huge mass of numbers is next to impossible even with the aid of an adding machine.

A more convenient and efficient method is to group the data into the form of a frequency distribution, and then compute the arithmetic mean for the distribution.

The arithmetic mean may be computed from the distribution in table 3 by assuming that all the values included within the limits of a class interval are distributed evenly in it[1] and therefore the average value of all the cases in each class interval coincides with the midpoint.

Long Method

Since no knowledge is available of the actual distribution of the cases within each group, it may be assumed that the cases are distributed evenly between the limits of the group. This would result in an average value for all values in the group equal to the midpoint of the group. Thus, the total value for each group may be obtained by multiplying the mid-point of the group by the number of cases in the group. (See table 3).

Thus for the frequency distribution in table 3 the midpoint

Table 3 — Computation of Arithmetic Mean by Long Method — Grouped Data
City Tax Rate on "True" Valuation in 261 Cities in the United States, 1927

(1) Rates Per Thousand Dollars (In Dollars) Class Interval	(2) Midpoint (*M. P.*)	(3) Number of Cities frequency (*f*)	(4) frequency × Midpoint (*f*) × (*M. P.*)
$4– 7.99	$6	5	30
8–11.99	10	15	150
12–15.99	14	46	644
16–19.99	18	68	1224
20–23.99	22	58	1276
24–27.99	26	32	832
28–31.99	30	22	660
32–35.99	34	10	340
36–39.99	38	2	76
40–43.99	42	2	84
44–47.99	46	0	0
48–51.99	50	1	50
		261	5366

Source: United States Department of Commerce, *Financial Statistics of Cities*, 1927, Table 23.

[1] If the class interval is not too large and a sufficient number of cases are available, variation from the assumption will be small and an accurate result will be obtained.

of the first class interval ($6) is multiplied by the frequency indicated for that group (5) in order to obtain the total value for all cases in the class interval. The totals (table 3, column 4) are then added to obtain the total value of all cases in the frequency distribution. The sum is divided by the number of cases (N) to obtain the arithmetic mean.

This method may be generalized into formula form as:

$$\bar{X} = \frac{\Sigma (f \times M.P.)}{N} = \frac{5366}{261} = \$20.56$$

The above is known as the long method because of the complex calculations which may result when the frequencies and the midpoint values are large.

Short Method

a. *Unit Deviation Method*

A simpler method may be devised by an examination of the characteristics of the arithmetic mean. If the mean is computed for a number of individual items (as shown below) and the deviation (distance) of each of the items from the mean is obtained, these deviations will total up to zero.[1]

Grades of Ten Students on an Examination in Arithmetic

Student Number	Grade (Per Cent)	Deviation from Mean (x)
1	95 %	15 %
2	92	12 %
3	90	10 %
4	86	6 %
5	86	6 %
6	80	0 %
7	75	− 5 %
8	72	− 8 %
9	64	− 16 %
10	60	− 20 %
	Total 800 %	0

$$\text{Mean } (X) = \frac{800\%}{10} = 80\%$$

If, however, some point other than the true arithmetic mean is selected the sum of the deviations from this point will not be zero. In the same series, for instance, 90% may be selected as an arbitrary starting point known technically as the **Guessed Mean**, and identified by the symbol $\bar{Z}$.

[1] The letter x is assigned as the symbol for deviation from the arithmetic mean

Student Number	Grade (Per Cent)	Deviation from Guessed Mean (d)
1	95 %	+ 5 %
2	92	+ 2 %
3	90	0 %
4	86	− 4 %
5	86	− 4 %
6	80	− 10 %
7	75	− 15 %
8	72	− 18 %
9	64	− 26 %
10	60	− 30 %
	800 %	− 100 %

If the *average* deviation of each item from the guessed mean is obtained:

$$\frac{\Sigma\,(d)}{N} = \frac{-100\%}{10} = -10\% \text{ (average deviation)}$$

and if this value is added to the arbitrary starting point ($\overline{Z}$) the result will be the arithmetic mean.[1]

$$\overline{X} = \overline{Z} + \frac{\Sigma\,(d)}{N} = 90\% + \frac{(-100\%)}{10} = 80\%$$

Where

$\overline{Z}$ = guessed mean
d = deviation of each value from guessed mean
N = number of cases

Table 4 — Computation of Arithmetic Mean — Short-Unit Deviation Method
Ratio of Current Assets to Current Liabilities for 221 Industria Corporations in the United States, 1930

(1) Ratios (Class Interval)	(2) (Midpoint) ($M.P.$)	(3) Number of Companies (frequency) (f)	(4) (deviation) (d)	(5) (frequency × deviation) fd
0– 1.99	1	11	− 4	− 44
2– 3.99	3	53	− 2	− 106
4– 5.99	5	47	0	0
6– 7.99	7	37	2	74
8– 9.99	9	21	4	84
10–11.99	11	16	6	96
12–13.99	13	13	8	104
14–15.99	15	8	10	80
16–17.99	17	10	12	120
18–19.99	19	1	14	14
20–21.99	21	2	16	32
22–23.99	23	1	18	18
24–25.99	25	0	20	0
26–27.99	27	1	22	22
		221		494

Source: Moody's Investors Service, *Moody's Industrials*, 1931.

[1] See technical appendix 1 for mathematical proof.

This technique may readily be applied to grouped data. For the distribution above (table 4) an *arbitrary starting point* (guessed mean) may be selected. Though any value may be taken for convenience the midpoint of one of the class intervals is generally used. Here 5.00 (midpoint of third class interval, table 4) may be used as the guessed mean.

The deviation from the arbitrary mean of the items in each group may then be obtained by getting the difference between the midpoint of each class interval and the guessed mean. Since the midpoint of the group is the average value of all items in the group this value (d) will represent the *average deviation* of the items in the group from the assumed mean. To obtain the total deviation for all items in the class interval it is necessary to multiply this deviation (d) by the frequency of the group (f). This is totaled for all class intervals to obtain the total deviation from the guessed mean of all values, and is then divided by N to obtain the average deviation about the guessed mean resulting in:

$$\frac{\Sigma\,(fd)}{N} = \frac{494}{221} = 2\,24$$

The above value is now added to the arbitrary starting point (guessed mean) to obtain the true arithmetic mean:[1]

$$\overline{X} = \overline{Z} + \frac{\Sigma\,(fd)}{N}$$

$$\overline{X} = 5.00 + \frac{494}{221} = 7.24$$

Where

$\overline{Z}$ = guessed mean
f = frequency of each class interval
d = deviation of midpoint of each group from guessed mean
N = total number of cases

b. Group Deviation Method

The computation of the arithmetic mean from a frequency distribution may be further simplified after consideration of the characteristics of such a distribution.

If the distribution has class intervals of a uniform size[2] it will be noted that the deviation of the midpoint of one group from the next will always be constant and equal to the size of the class interval. In the distribution shown in table 4 there is a constant difference of 2 between the midpoints, and this is equal to the size of the class intervals—for example 0 to 1.99.

The deviations of any one group from any other group may then be measured in terms of class intervals. In the distribution below

[1] See technical appendix I for mathematical derivation of this formula.
[2] Wherever possible this should be the rule in compiling such a distribution

(table 5) the midpoint of the 3rd class interval (5.00) has again been selected as the arbitrary or guessed mean ($\overline{Z}$). The midpoint of the first class interval deviates from the guessed mean by the amount of -4, or -2 class intervals. The deviation column (table 5, column 4) is expressed in terms of class intervals rather than in the original units of the data. The resulting values in the deviation column are numerically smaller and thus the computation is simpler.

Table 5 — Computation of Arithmetic Mean — Short-Group Deviation Method Ratio of Current Assets to Current Liabilities for 221 Industrial Corporations in the United States, 1930

(1) Ratio (Class Interval)	(2) (Midpoint) ($M.P.$)	(3) Number of (Companies Frequency) (f)	(4) (Deviation In Class Intervals) (d')	(5) (fd')
0– 1.99	1	11	– 2	– 22
2– 3.99	3	53	– 1	– 53
4– 5.99	5	47	0	0
6– 7.99	7	37	1	37
8– 9.99	9	21	2	42
10–11.99	11	16	3	48
12–13.99	13	13	4	52
14–15.99	15	8	5	40
16–17.99	17	10	6	60
18–19.99	19	1	7	7
20–21.99	21	2	8	16
22–23.99	23	1	9	9
24–25.99	25	0	10	0
26–27.99	27	1	11	11
		221		247

Source: Moody's Investors Service, *Moody's Industrials*, 1931.

$$\overline{X} = \overline{Z} + \frac{\Sigma (fd')}{N} \qquad = 5 + \frac{247}{221}(2) = 7.24$$

As in the previous method the computation is then carried out in the same manner and results in an average deviation about the guessed mean, $\frac{\Sigma (fd')}{N}$, now in terms of class intervals. To convert this back to original values it is multiplied by the size of the class interval. The result may then be added to the guessed mean to obtain the arithmetic mean.

This method may be generalized into formula form as follows:

$$\overline{X} = \overline{Z} + \frac{\Sigma (fd')}{N} C$$

Where[1]

$\overline{Z}$ = Guessed Mean

f = frequency

This formula is also variously given by different texts as:

$$M = A + \frac{1}{N} \Sigma (f) C \qquad A = E + \frac{\Sigma f (V - E)}{} C \qquad M = M' + c$$

The Mode

Definition

The mode is the most frequent or most common value, provided that a sufficiently large number of items are available to give a smooth distribution.

The value of the mode will correspond to the value of the maximum point (ordinate) of a frequency distribution if it is an "ideal" or smooth distribution.

Computation

It is not possible to make an exact mathematical determination of the mode. A number of methods may be used, however, to secure reasonably accurate approximations.

The midpoint of the modal class interval may not be used as the value of the mode, since its value will change if the size of the class interval is changed.

Reducing the size of the class interval will tend to delimit the value of the mode and tend more and more to have it coincide with the midpoint of the group of greatest frequency. This reduction in size of the class interval is, however, decidedly limited by the number of items included in the distribution. If an infinite number of items are available and an infinitely small class interval is used, the midpoint of the class interval of greatest frequency will be the value of the mode.

In practice this ideal situation does not exist, so that an approximation somewhat closer than the midpoint of the modal group is necessary.

In spite of the previous midpoint assumption the values within a group are not evenly distributed in a distribution, but there is a tendency to gravitate towards the point of greatest density.

In the distribution below (table 7) the modal group is that containing 43 items, with class limits of .10 % to .19 %. Since there are a greater number of cases in the class above, .20 % to .29 %, with a frequency of 32, than that below[1], .00 % to .09 %, which contains 19 items, it follows that the true point of greatest concentration will tend towards the upper class interval and will therefore be above the midpoint of the modal group.

The value of the mode may be approximated by resort to the formula:

$$\text{Mode} = L_{mo} + \frac{f_a}{f_a + f_b} C = .10\% + \frac{32}{32 + 19}(.10\%) = .163\%$$

where

L_{mo} = lower limit of modal group
f_a = frequency of class interval above modal group
f_b = frequency of class interval below modal group
C = size of class interval

[1] By *below* in reference to a class interval is meant in the direction of the lowest class interval value.

Table 7 — Computation of Mode Moments of Force Method
Percent of Population on Old Age Pensions in the United States, by Counties, 1930

(Class Intervals) Percent	Number of Counties (Frequency)
.00- .09%	19
.10- .19	43
.20- .29	32
.30- .39	27
.40- .49	17
.50- .59	21
.60- .69	14
.70- .79	9
.80- .89	2
.90- .99	2
1.00-1.09	0
1.10-1.19	0
1.20-1.29	1
	187

Source: United States Bureau of Labor Statistics, *Handbook of Labor Statistics*, Bulletin 451, 1931, pp. 483–487.

The above technique is sometimes referred to as the **moments of force** method.

Empirical Method

Where the distribution is only moderately skewed another estimation of the value of the mode may be obtained from the relationship that exists between the position of the mean, median and mode.

In a smoothed curve (such as that shown in figure 10) the mode will be located at the highest point in the distribution, the position of the median will be somewhat to the right of the mode in the direction of the extreme values and will divide the area under the curve in half. The mean, since it is affected to the greatest degree by extreme values, will be furthest in the direction of the extreme values in this hypothetical right skewed distribution.

It has been found that for a moderately skewed distribution **the distance between the mean and the median is one-third of the distance between the mean and the mode.**

In a left skewed distribution the same relationship will occur but in the opposite direction.

Since the values of the mean and the median may be determined exactly, the value of the mode may be approximately determined through this relationship.

$$\text{Mode} = \text{Mean} - 3\,(\text{Mean} - \text{Median})$$

Other Methods

Estimates of the value of the mode may be determined by a number of additional methods such as:[1]

[1] More advanced methods of determining the value of the mode are outlined in Chapter XIV

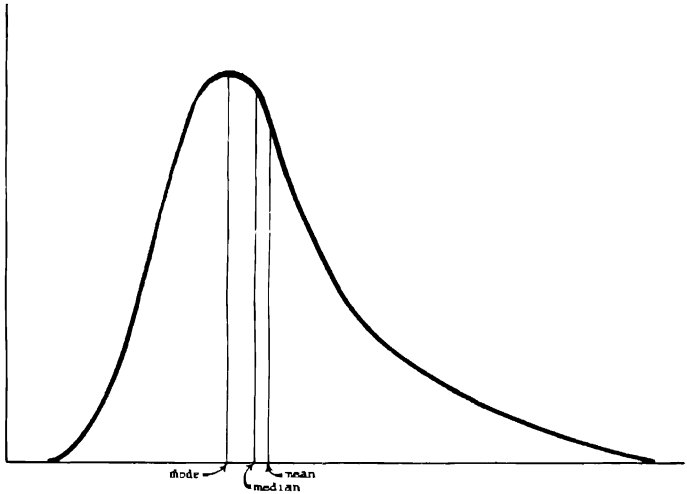


Fig. 10—Hypothetical right skewed frequency distribution showing theoretical position of mode, median and mean.

1. The grouping method.
2. By smoothing the frequency distribution
3. By moving averages.
4. By mathematical curves

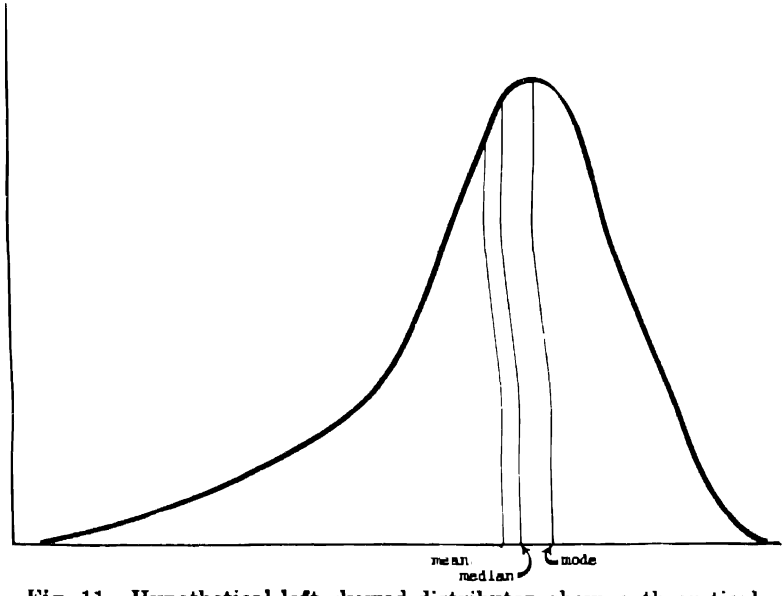


Fig. 11—Hypothetical left skewed distributon showng theoretical position of mode, median and mean.

Mode

Characteristics

1. By definition the mode is the most usual or typical value. Under certain circumstances it may be considered as the "normal" value.

2. The value of the mode is entirely independent of extreme items.

3. The mode is an average of position.

Advantages

1. It is the most typical and therefore the most descriptive average.

2. It is simple to approximate by observation where there are a small number of cases.

3. It is not necessary to arrange the values or know them if they are few in number.

Disadvantages

1. The mode can be approximated *only* when a limited amount of data is available.

2. Its significance is limited when a large number of values is not available.

3. In a small number of items the mode may not exist, for none of the values may be repeated.

The Geometric Mean

The geometric mean is the nth root of the product of n items or values.

Formula:

$$G_m = \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \ldots\ldots X_n}$$

For instance the geometric mean of \$1, \$3 and \$9 is

$$G_m = \sqrt[3]{1 \times 3 \times}$$
$$= \sqrt[3]{27}$$

To facilitate the computation of the geometric mean, its formula may be reduced to its logarithmic form.

$$\log G_m = \frac{\log X_1 + \log X_2 + \log X_3 + \ldots\ldots + \log X_n}{N}$$

where

G_m = Geometric Mean

It can be seen that the logarithm of the geometric mean is equal to the average of the logarithms of the items.

The geometric mean may be computed from grouped data by using the technique outlined for the computation of the arithmetic mean by the "long" method (see page 12), except that the logarithms of the midpoints are used in the calculations rather than the actual midpoint values.

Characteristics

1. The geometric mean is a calculated value and dependent upon the size of all the values.
2. It is less affected by extreme items than the arithmetic mean.
3. For any series of items it is always smaller than the arithmetic mean.

Advantages

1. It is a more typical average than the arithmetic mean since it is less affected by extremes.
2. It may be manipulated algebraically.
3. It is particularly useful in the computation of index numbers (see Chapter XIII).

Disadvantages

1. The geometric mean is not widely known.
2. The geometric mean is relatively difficult to compute.
3. It cannot be determined where there are negative values in the series or where one of the items is zero.

The Quadratic Mean

The quadratic mean is the square root of the mean square of the items (root-mean-square).

Formula:

$$Q_m = \sqrt{\frac{\Sigma (X^2)}{N}}$$

The quadratic mean is used in the computation of the standard deviation (see page 34).

The Harmonic Mean

The harmonic mean of a series of values is the reciprocal of the arithmetic mean of the reciprocals of the values.

Formula:

$$\frac{1}{H_m} = \frac{\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} \quad \ldots\ldots\ldots + \frac{1}{X_n}}{N}$$

The harmonic mean is used in averaging rates.

ADDITIONAL BIBLIOGRAPHY*

BOWLEY, ARTHUR L., *Elements of Statistics*, pp. 86–109. P. S. King & Son, London, 1907.

KELLEY, TRUMAN L., *Interpretation of Educational Measurements*, pp. 185–188. World Book Co., Yonkers, New York, & Chicago, Illinois.

ODELL, C. W., *Educational Statistics*, pp. 147–179. Century Co., New York, 1925.

OTIS, ARTHUR S., *Statistical Method in Educational Measurements*, pp. 11–19. World Book Co., Yonkers, New York, & Chicago, Illinois.

RIETZ, H. L. (Editor), *Handbook of Mathematical Statistics*, pp. 25–27. Houghton Miffln Co., New York, 1924.

SUTCLIFFE, WILLIAM G., *Statistics for the Business Man*, pp. 69–76. Harper & Bros., New York, 1930.

TRAUBE, MARION R., *Measuring Results in Education*, pp. 201–209; 216–225. American Book Co., New York, 1924.

ZIZEK, FRANK, *Statistical Averages*, pp. 194-247. Henry Holt & Co., New York, 1930.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents

CHAPTER IV

FREQUENCY DISTRIBUTION
DISPERSION AND SKEWNESS

Dispersion

The average or typical value is of little use unless the degree of variation which occurs about it is given.

If the scatter about the measure of central tendency is very large it is of little use as a typical value. It is therefore necessary to develop a quantitative measure of the dispersion (or variation or scatter) which occurs about the average.

The Range

The range, the simplest of the measures of dispersion, is the difference between the minimum and maximum items in the series. It is sometimes given in the form of a statement of the minimum and maximum values themselves.

The difference between these two values gives some idea of the degree of variation occurring in the series, but quite frequently the result is misleading.

In series A and B below *the range* is 30%, but the dispersion is not the same.

Hypothetical Examination Grades for Ten Students

	Grade in Percent	
Student Number	Examination A	Examination B
1	60 %	60 %
2	60 %	65 %
3	61 %	70 %
4	63 %	72 %
5	65 %	75 %
6	65 %	78 %
7	66 %	80 %
8	67 %	85 %
9	68 %	88 %
10	90 %	90 %

Characteristics

1. The range is simple and readily understood.
2. It is easily calculated.
3. Its value is dependent on two items only, the highest and lowest values.
4. The distribution of the items between the two extremes is not necessary to obtain the range.
5. Since the range is dependent only upon the two extremes, it is greatly affected by unusual occurrences when these two items are distinctly "out of line."

Mean Deviation

The range is dependent for its value entirely upon the two extremes. Obviously, where there are extremes a satisfactory measure of dispersion must be dependent upon the position of every value in the series.

A simple method for determining the scatter of a series of values about a given point (i.e., the scatter of hits on a target) would be to take the average distance of the items from the given point (with the target this point would be the "bull's eye"). The smaller the average distance about this point the smaller the scatter or dispersion of the values. In a frequency distribution the average distance of the items from the measure of central tendency, such as the arithmetic mean, may be used for this purpose.

Since, however, the sum of the deviations about the arithmetic mean total up to zero, in order to obtain the average value it is necessary to ignore signs.

Table 8—Sales Record of Clerk No. 148 in a New York City Department Store for the Month of June 1932

Date	Number of Sales (X)	Deviation from Monthly Average (x)
June 1	15	11.85
2	27	.15
3	31	4.15
4	27	.15
6	23	3.85
7	23	3.85
8	25	1.85
9	31	4.15
10	29	2.15
11	41	15.15
13	17	9.85
14	30	3.15
15	45	18.15
16	24	2.85
17	26	.85
18	26	.85
20	23	3.85
21	15	11.85
22	37	10.15
23	27	.15
24	18	8.85
25	39	12.15
27	19	7.85
28	18	8.85
29	21	5.85
30	41	13.15
Total	698	165.70
Average	26.85	6.37

This measure of dispersion is called the **mean deviation**. It consists of the average of the deviations of the items from their arithmetic mean or median.

Mean Deviation

Characteristics

1. The value of the mean deviation is dependent upon the value of every item in the series.
2. It may be computed about either the arithmetic mean or the median.
3. The average deviation from the median is a minimum.

Computation Mean Deviation—Ungrouped Data

Formula:

$$MD = \frac{\Sigma |x|}{N} \text{ or } \frac{\Sigma |d|}{N}$$

Where

MD is the mean deviation

$\Sigma |x|$ = a sum of the deviations of each value from the arithmetic mean, *signs ignored.*

$\Sigma |d|$ = the sum of the deviations from another measure of central tendency such as the median, *signs ignored.*

The mean deviation may be computed about either the mean or the median. When computed about the median it will be smaller than about the mean or, since the sum of the deviations about the median is a minimum, about any other value.

Computation—Grouped Data

When the data is grouped in the form of a frequency distribution the value of the mean deviation may be determined as follows:

1. Obtain the deviation of the midpoint of each class interval from the median (or mean).
2. Multiply the deviations by the number of items (the frequency) in each class interval.
3. Divide the total of the values obtained by the number of cases.

A simpler method (arithmetically) is to:

1. Select an arbitrary origin.
2. Obtain the deviations of the midpoints about this more convenient value. The midpoint of the group in which the median (or mean) is located is selected as an origin. The deviations of the midpoint of each value is then determined (see d' in table 9, column 4) and multiplied by the frequency of each group (column 3).

In the illustrated distribution the median (73.052 months) is located above the midpoint of the class interval used as the arbitrary origin, (73); and therefore all of the deviations at that point

Table 9—Computation of Mean Deviation

Ages of Pupils in the First Half of First Grade in a New York City Public School

(1) Age in Months Class Intervals	(2) Midpoint ($M\ P$)	(3) Number of Pupils (F)	(4) Deviation from $\bar{Z}$ (d') (In Class Intervals)	(5) Frequency × Deviation (fd')
68–69 9	69	12	2	24
70–71 9	71	33	1	33
72–73 9	73	57	0	0
74–75.9	75	25	1	25
76–77 9	77	9	2	18
78–79.9	79	4	3	12
80–81 9	81	6	4	24
82–83.9	83	2	5	10
84–85 9	85	0	6	0
86–87 9	87	0	7	0
88–89 9	89	2	8	16
		150		162

or below are too small by the amount of the difference between the two values, .052 months in actual units or .026 class intervals. There are 102 (12 + 33 + 57) of these values and therefore their total understatement is 102 times the differences of .026 class intervals.

The values above the arbitrary origin in similar fashion are overstated by the amount of the difference times the number of the values or .026 class intervals times 48.

The understatement of the 102 values below the arbitrary origin is in part offset by the 48 items overstated, leaving only 54 values understated. The average understatement will then be 54 times .026 divided by the total number of items. If this correction is added to the average deviation about the arbitrary origin the result will be the mean deviation in class intervals. Formula[1]

$$MD' = \frac{\Sigma\,|fd'|}{N} + \frac{(N_S - N_L)\,c}{N} = \frac{162}{150} + \frac{(102-48)\ .026}{150} = 1.0894$$

[1] (a) The equation may be simplified to read

$$M\ D = \frac{\Sigma\,(fd) + (N_s - N_L)c}{N}\,C$$

(b) This calculation assumes all values to be located at the midpoint in the median or mean group A more exact value may be obtained by assuming an even distribution of these values throughout that group The following formula (see *Handbook of Mathematical Statistics* H. L Rietz, Editor pp 29–31) may be used

$$MD' = \frac{\Sigma fd}{N} + \frac{(N_b - N_a)c + f_m\,(25 + c^2)}{N}$$

Where N_a = number of items above mean (or mean) group
N_b = number of items below median (or mean) group
f_m = number (frequency) of cases in median (or mean) group
c = differences between arbitrary origin and median (or mean)

MD in actual values is obtained by multiplying MD' by C

Where

MD' = Mean deviation in class intervals
N_S = Number of cases too small or understated
N_L = Number of cases too large or overstated
c = Difference between arbitrary origin (midpoint of median or mean group) and median or mean

The mean deviation in actual units will be obtained if the result, which is expressed in terms of class intervals, is multiplied by the size of the class interval.

$$MD = MD' \times C$$

where

C = size of class interval

$MD = 1.0894 \times 2 = 2.1788$ months

The Standard Deviation

The standard deviation is a special form of average deviation from the mean. It is computed by taking the quadratic mean (see page 27) of the deviations from the arithmetic mean of these values. The standard deviation is thus the root-mean-square of the deviations from the arithmetic mean.

Formula:

$$\sigma = \sqrt{\frac{\Sigma (x^2)}{N}}$$

where

σ = standard deviation*
x = deviations from arithmetic mean
N = total number of items, $\Sigma (f)$

Computation—Ungrouped Data[1]

1. Get the difference between each actual value and the arithmetic mean.
2. Square the values thus obtained. Obtain the average of the squares.
3. Take the square root of the resulting total.

Computation—Grouped Data

Where there are a considerable number of items in the series the calculation of the standard deviation can be more readily performed if the data are first grouped into the form of a frequency distribution.

1. The deviation of the midpoint of each group from the arithmetic mean is used as a measure of the average deviation from the mean of all items in that group.

* The symbol σ is the Greek small letter Sigma.

[1] A more convenient formula for ungrouped data may be derived from an algebraic manipulation of the standard deviation: $\sigma = \sqrt{\frac{\Sigma (X^2)}{N} - \left(\frac{\Sigma X}{N}\right)^2}$

See technical appendix III.

Table 10—Computation of Standard Deviation—Ungrouped Data
Bid Prices for 12 Joint Stock Land Bank Bonds on April 18, 1934

Bank	Rate (Percent)	Bid Price (X)	Deviation from Mean (70 5) ($X - \overline{X}$) x	x^2
Atlanta	5%	71	0.5	.25
Burlington	5	65	− 5.5	30.25
Chicago	5	41	− 29 5	870.25
Dallas	5	80	9.5	90.25
Denver	5	73	2.5	6.25
Des Moines	5	78	7.5	56.25
Fort Wayne	5	71	.5	.25
First Carolinas	5	69	− 1.5	2 25
First Texas	5	71	.5	.25
Lincoln	5	79	8.5	72.25
Louisville	5	75	4.5	20.25
New York	5	73	2.5	6.25
Total		846	0	1155.00
Mean		70 5		96.25

Source: *Wall Street Journal.*

$$\sigma = \sqrt{\frac{\Sigma}{N}(x^2)} \qquad 9.81$$

Table 11—Computation of Standard Deviation—Grouped Data—Long Method
Percent of Tax Delinquency in 151 Cities of Over 50,000 Population in the United States, 1933

(1) Percent of Tax Delinquency (Class Interval)	(2) (Midpoint) $M.P.$	(3) Number of Cities (Frequency) (f)	(4) Deviation from Mean (26 74) (x)	(5) (x^2)	(6) $f(x^2)$
0– 4.99	2 50	1	− 24 24	587 5776	587 5776
5– 9 99	7.50	12	− 19 24	370 1776	4442 1312
10–14.99	12.50	19	− 14 24	202 7776	3852 7744
15–19 99	17 50	24	− 9 24	85 3776	2049 0624
20–24.99	22 50	19	− 4 24	17 9776	341 5744
25–29.99	27.50	19	.76	.5776	10 9774
30–34.99	32 50	16	5 76	33 1776	530 8416
35–39.99	37.50	15	10 76	115 7776	1736 6640
40–44.99	42 50	12	15 76	248 3776	2980 5312
45–49.99	47.50	8	20 76	430.9776	3447 8208
50–54.99	52.50	2	25.76	663 5776	1327 1552
55–59.99	57.50	0	30.76	946 1776	0
60–64.99	62.50	2	35 76	1278.7776	2557 5552
65–69.99	67.50	2	40.76	1661.3776	3322 7552
		151			27187 4176

Source: Dun & Bradstreet's Municipal Review.

$$= \sqrt{\frac{\Sigma f(x^2)}{N}} = \sqrt{\frac{27{,}187.4176}{151}} = 13.42\%$$

2. The average deviation of each group is squared to obtain the necessary deviation squared.
3. The average deviation is multiplied by the frequency indicated for the group in order to obtain the total of the squared deviations for that group.
4. The totals are then added for the entire distribution.
5. The square root of the sum obtained after dividing by N is the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma f(x^2)}{N}}$$

Short Method

The computation of the standard deviation may be simplified by the following method.

1. Instead of using the arithmetic mean, compute the standard deviation about a conveniently selected point. For this purpose the midpoint of any group may be selected. Since the quadratic mean of the deviations about the arithmetic mean is smaller than the quadratic mean of deviations about any other point, the resulting value will be larger than the true standard deviation.
2. Subtract from it a correction value to obtain the necessary result. The value of the correction factor may be determined by an algebraic manipulation of formula $\sqrt{\frac{\Sigma f(x^2)}{N}}$ (see technical appendix 11).

The resulting formula is

$$= \sqrt{\frac{\Sigma f(d^2)}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

where σ = standard deviation
d = deviation of midpoint of each class interval from that of arbitrary group.

Where the class intervals are uniform in size the calculation may be further simplified by carrying out all computations in terms of class intervals and then multiplying the final results by the size of the class interval.

The formula will then read

$$\sigma = C\sqrt{\frac{\Sigma f(d')^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2}$$

where σ = standard deviation
d' = deviation of midpoint of class interval from arbitrary origin in terms of class intervals.
f = frequency of values in class interval.
C = size of class interval.

The application of this formula to the computation of the standard deviation of the minimum line rates for national advertising for 249 daily newspapers is shown in table 12.

Table 12—Computation of Standard Deviation—Short-Group Deviation Method Minimum Line Rate for National Advertising for 249 Daily Newspapers in Cities of 25,000 to 50,000 in the United States, 1933

(1) Rate Per Line (In Dollars)	(2) Number of Newspapers (f)	(3) Deviation from $\bar{Z}$ In Class Intervals (d')	fd'	fd'^2
$ 01- 019	2	- 5	- 10	50
.02- 029	4	- 4	- 16	64
.03-.039	23	- 3	- 69	207
.04-.049	30	- 2	- 60	120
.05-.059	40	- 1	- 40	40
.06-.069	45	0	0	0
.07- 079	35	1	35	35
.08-.089	25	2	50	100
09-.099	12	3	36	108
.10- 109	9	4	36	144
11- 119	6	5	30	150
.12-.129	10	6	60	360
.13-.139	3	7	21	147
.14-.149	1	8	8	64
.15-.159	1	9	9	81
.16-.169	3	10	30	300
	249		120	1970

Source: Editor and Publisher, *International Year Book for 1933.*

$$\sigma = C\sqrt{\frac{\Sigma f(d'^2)}{N} - \left(\frac{\Sigma fd'}{N}\right)^2}$$

$$= \$.01\sqrt{\frac{1970}{249} - \left(\frac{120}{249}\right)^2} \qquad = \$.01\sqrt{7.9116 - .2322}$$

$$= \$.01\sqrt{(7.6794)} \qquad = (2.7711)\ \$.01 = \$.0277$$

Correction for Grouping

The value computed by the grouping method will be subject to the assumption that all the values are located at the midpoint of each class interval and in part is dependent on the size of the class interval used. The error in the assumption is constant when:

1. The distribution is "continuous" (see page 7).
2. The distribution tapers off gradually in both directions. Under these conditions the true standard deviation is $\sigma^2 = (\sigma'^2 - 1/12)\ C^2$

where:

σ' = standard deviation in class interval units
C = size of class interval

Check on Computation—Charlier Check

A simple check may be used in order to determine the accuracy of the computations preliminary to the actual substitution in the formula for the standard deviation.

If an additional value $\Sigma f (d' + 1)^2$ is computed it will be seen that

$$\Sigma f (d' + 1)^2 = \Sigma f (d'^2 + 2d' + 1)$$
$$= \Sigma f (d'^2) + 2 \Sigma (fd') + \Sigma (f)$$

But since

$$\Sigma f = N$$
$$\therefore \Sigma f (d' + 1)^2 = \Sigma fd'^2 + 2 \Sigma (fd') + N$$

These are the values used in the computation of the standard deviation. If the equation is fulfilled the preliminary values are correct. If the computation check is applied to the problem shown above the result is:

Table 12a—Charlier Check—Applied to Data of Table 12

(1) Rate Per Line (In Dollars)	(2) Number of Newspapers (f)	(3) (From Table 12) d'	(4) $d' + 1$	(5) $(d' + 1)^2$	(6) $f(d' + 1)^2$
$.01–.019	2	– 5	– 4	16	32
.02–.029	4	– 4	– 3	9	36
.03–.039	23	– 3	– 2	4	92
.04–.049	30	– 2	– 1	1	30
.05– 059	40	– 1	0	0	0
.06–.069	45	0	1	1	45
.07–.079	35	1	2	4	140
.08– 089	25	2	3	9	225
.09–.099	12	3	4	16	192
10–.109	9	4	5	25	225
.11–.119	6	5	6	36	216
.12–.129	10	6	7	49	490
.13–.139	3	7	8	64	192
.14–.149	1	8	9	81	81
.15–.159	1	9	10	100	100
.16–.169	3	10	11	121	363
	249				2459

$$\Sigma f (d' + 1)^2 = 2459 =$$
$$\Sigma fd'^2 + 2 \Sigma fd' + N = 1970 + 2 (120) + 249$$

Characteristics

1. The standard deviation is affected by the value of every item.

2. Greater emphasis is placed on extremes than in the mean deviation, this because all values are squared in the computation of the standard deviation.

3. In a normal or bell shaped distribution the mean deviation

is .7979 σ. In a moderately skewed distribution this relationship is approximately true.

4. (a) If a distance equal to one standard deviation is measured off on the X axis on both sides of the arithmetic mean in a normal distribution, 68.26% of the values will be included within the limits indicated.
   (b) If two standard deviations are measured off 95.46% of the items will be included.
   (c) Three standard deviations measured off will include 99.73% of the cases.

The above percentages are exact only where the distribution is perfectly normal. In the case of a moderately skewed distribution the percentages are approximations. As such they are indicated more generally as about 68% for one standard deviation on either side of the mean, 95% for two, and practically all of the values (99.7%) for three.

The exact per cent of cases included for any number of standard deviations measured from the arithmetic mean in one direction only may be found in table 31, page 110.

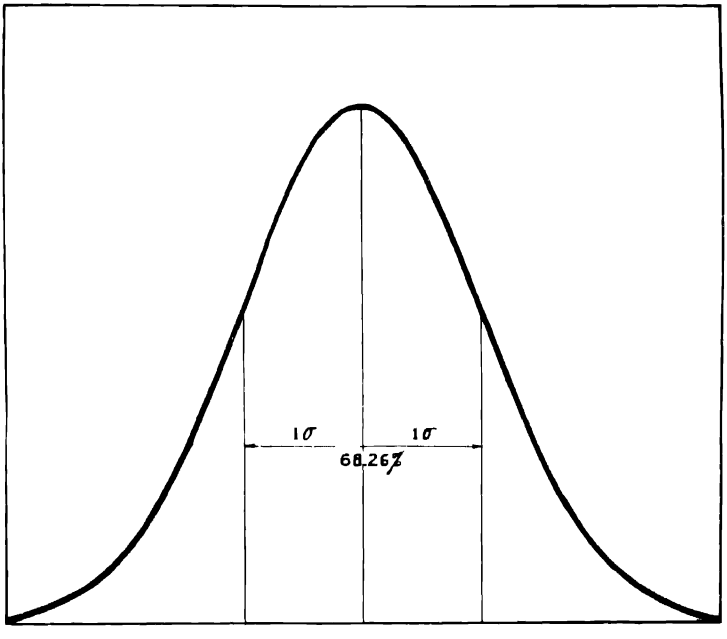


Fig. 12—"Normal" Distribution Showing the Percentage of the Area Included Within One Standard Deviation Measured Both Plus and Minus About the Arithmetic Mean.

5. Since 3 standard deviations on either side of the mean include all (99.7%) of the cases in moderately skewed or normal distributions, the standard deviation is about $\frac{1}{6}$ of the range.

The Quartile Deviation or Semi-Interquartile Range

As the dispersion of a frequency distribution is increased the distance between the quartiles is enlarged. Since an increased dispersion will be indicated by a greater distance between the quartiles, this distance may be used as the basis of a measure of dispersion.

If the distribution were perfectly symmetrical the two quartiles would be equi-distant from the median. One half of the distance between the quartiles would represent the distance between the quartiles and the median.

One half of the distance between the quartiles may be used as a measure of the average distance of each quartile from the median. This value is used as a measure of dispersion.

$$QD = \frac{Q_3 - Q_1}{2}$$

where:

QD = Quartile Deviation
Q_3 = Third Quartile
Q_1 = First Quartile

If a distance equal to one QD is measured off on either side of a point half way between the quartiles, 50% of the values will be included between these limits. This midpoint value is assigned the letter K. K coincides with the median only in a symmetrical distribution.

The 10-90 Percentile Range

In a similar manner as the dispersion of the distribution is increased, the distance between the various percentiles will be increased. The distance between any two percentiles may thus be used as a measure of dispersion. The spread between the tenth percentile and ninetieth percentile is generally used for this purpose. This measure is known as the 10–90 percentile range

$$10 - 90 \text{ Percentile Range} = P_{90} - P_{10}$$

This measure of dispersion has the advantage of being dependent upon a larger percent (90%) of the cases than the quartile deviation (50%) while excluding the unusual cases represented by the extreme 10% on either end of the distribution.

Relative Measures of Dispersion

The measures outlined above are *absolute* measures of dispersion and therefore the resulting values cannot always be compared with significance. The standard deviation in months of the ages of

students in a given grade of a New York City school cannot be compared to the dispersion of the Intelligence Quotients in the same class where there is a standard deviation in percent because of the difference in units.

In addition a measure of dispersion must be compared to the size of the average about which it is measured. For instance, a variation of five dollars in the price of a share of stock, the average price of which is ten dollars, does not have the same value in relation to its average price as a share which has the same variation but an average price of one hundred dollars.

To relate the measure of dispersion to its average and to convert it to percentage form, the standard deviation is divided by the arithmetic mean. Stating this measure in percentage form solves the problem presented by the differing units. The resulting measure developed by Pearson is known as the **coefficient of variation** (V)

$$V = \frac{\sigma}{\overline{X}} 100$$

Other comparative coefficients of dispersion may be computed when using the other measures of dispersion.

$$V_{AD} = \frac{AD}{\text{Median (or mean)}}$$

$$V_Q = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Skewness

Skewness is a term for the degree of distortion from symmetry exhibited by a frequency distribution.

When a distribution is perfectly symmetrical, the values of the mean, median and mode coincide. In an asymmetrical (skewed) distribution the values of the averages will depart from one another. Since the arithmetic mean is most affected by extremes it will move the greatest distance from the mode. The mode is not affected at all by unusual values; therefore the greater the degree of skewness the greater the distance between the mean and the mode.

It follows that this distance between mean and mode may be used as a measure of skewness, since the greater the lack of symmetry the larger the discrepancy between them. However, because the measure of skewness is used largely for comparative purposes, the problem of differing units will again make its appearance. A second difficulty arises in that the distance between the averages will be larger in a widely dispersed distribution than in one with a narrow dispersion. Both difficulties may be removed by dividing the measure by the measure of dispersion.

$$S_K = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Since the distance between the mean and mode in moderately skewed distributions is three times the distance between the mean and the median (see page 24), for this type of distribution the formula may be rewritten as follows:

$$S_K = \frac{3\ (\text{Mean} - \text{Median})}{\sigma}$$

When the distribution is symmetrical the values of the mode and the mean will coincide. Under these circumstances the coefficient of skewness will be zero.

Where the curve is right skewed the extremely large values will *increase* the value of the mean. This results in increasing the value of mean over that of the mode, since the latter remains unaffected by the extremes. The coefficient will then be a positive value. If the distribution is skewed to the left, the extreme cases will *reduce* the value of the mean. This makes it smaller than the mode and results in a negative coefficient of skewness.

Another measure is based on the position of the quartiles. In a symmetrical distribution the quartiles are equidistant from the median. In a skewed distribution the quartiles will differ in their distances from the median.

The greater the lack of symmetry the larger the discrepancy between the two distances of the quartiles from the median. If this is divided by the quartile deviation (the measure of dispersion based on the quartiles) the result is a coefficient of skewness.

Formula:

$$S_K = \frac{(Q_3 - \text{Median}) - (\text{Median} - Q_1)}{QD}$$

Since the distances will be equal for a symmetrical distribution, in this case S_K will equal zero. Where the distribution is right skewed the right Quartile (Q_3) will be a greater distance from the median than Q_2. The opposite will be true where the curve is left skewed. The resulting coefficient will then be negative.

Kurtosis

The "peakedness" of the frequency distribution is another characteristic which might be measured. The measurement of this characteristic is outlined in Chapter XV.

ADDITIONAL BIBLIOGRAPHY*

BOWLEY, ARTHUR L., *Elements of Statistics*, pp. 110–116. P. S. King & Son, London, 1907.

KELLEY, TRUMAN L., *Interpretation of Educational Measurements*, pp. 51–54; 151–155. World Book Co., Yonkers, New York, & Chicago, Illinois, 1927.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

ODELL, C. W., *Educational Statistics*, pp. 117–145; 281–285. Century Co., New York, 1925.

OTIS, ARTHUR S., *Statistical Method in Educational Measurements*, pp. 85–94. World Book Co., Yonkers, New York & Chicago, Illinois, 1926.

REITZ, H. L., (Editor), *Handbook of Mathematical Statistics*, pp. 27–33. Houghton Miffin Co., New York, 1924.

TRAUBE, MARION R., *Measuring Results in Education*, pp. 252–273. American Book Co., New York, 1924.

ZIZEK, FRANK, *Statistical Averages*, pp. 251–338. Henry Holt & Co., New York, 1930.

CHAPTER V

TIME SERIES ANALYSIS—TREND

Definition

The time series is an arrangement of statistical data in accordance with its time of occurrence.

Classification of Movements

The analysis of the time series consists of the description and measurement of the various changes or movements as they appear in the series during a period of time. These changes or movements may be classified as.

1. **Secular trend,** or the long time growth or decline occurring within the data. The period covered should include not less than ten years.
2. **Seasonal variation,** or the more or less regular movement within the twelve month period. This movement occurs year after year and is caused by the changing seasons.
3. **Cyclical movement,** or the swing from prosperity through recession, depression, recovery, and back again to prosperity. This movement varies in time, length and intensity.
4. **Residual, accidental or random** variations, including such unusual disturbances as wars, disasters, strikes, fads, or other non-recurring factors.

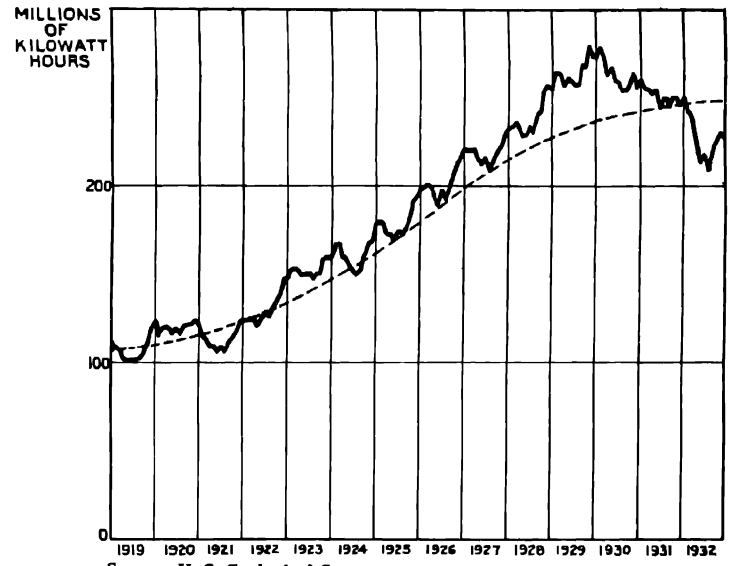
Measurement of Trend

Four methods are commonly used for measuring trends;

1. Freehand.
2. Semi-average.
3. Moving average.
4. "Least squares" (see Chapter VI).

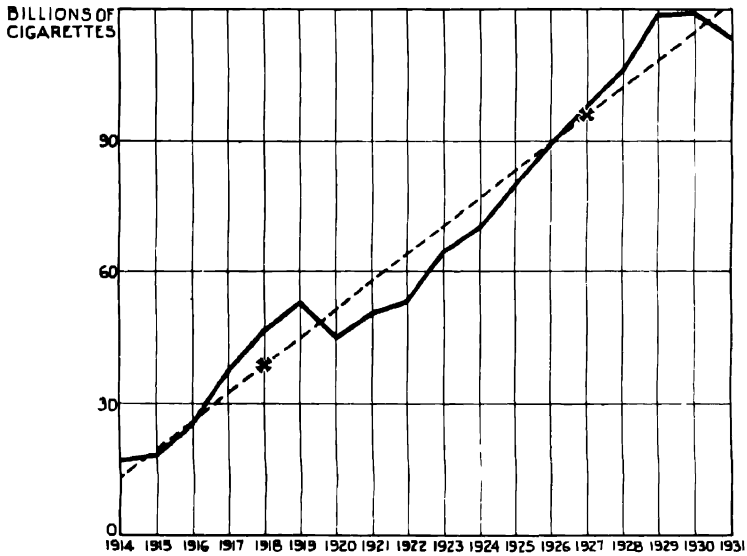
Methods

1. **Freehand Method.** To fit a trend by the freehand method draw a line through a graph of the data in such a way as to describe what appears to the eye to be the long period movement. A line of trend fitted by this method is shown in figure 13. The drawing of this line need not be strictly freehand but may be accomplished with the aid of a transparent straight edge or a "French" curve.



Source: U. S. Geological Survey.

Fig. 13—Average Daily Output of Electric Power in the United States, 1919—1932. Trend indicated by a freehand line.



Source: United States Department of Commerce.

Fig. 14—Consumption of Cigarettes in the United States, 1914—1931. Trend indicated by semi-average method.

Advantages

1. The method is simple.
2. The method may be used in place of a mathemetical equation which may not logically describe the trend.
3. If drawn with care the trend line fitted by this method will be a close approximation to a mathematically fitted trend.

Disadvantages

1. The results vary according to personal estimate.
2. Considerable practice is required to make a good fit.

2. **Semi-average Method.** In this procedure the data are split into two equal parts and the figures in each half are averaged. The averages thus obtained are plotted at the center of their respective periods (see figure 14) and a straight line is then drawn through the two points.

In the example shown below (table 13) the data are split into two parts (1914–1922 and 1923–1931). The two values obtained by averaging the figures in each half of the data (38.25 for the first half and 95.45 for the second half) are plotted at the midpoints of their respective periods (the middle of 1918 for the first group and the middle of 1927 for the second). With the aid of a ruler a straight line is then drawn through the two points.

Table 13—Computation of Trend—Semi-Average Method
Consumption of Cigarettes in the United States, 1914-1931

Year	Consumption (Billions of Cigarettes)	Totals	Arithmetic Means
1914	16 86		
1915	17 96		
1916	25 29		
1917	35 33		
1918	46.66	344 28 + 9 =	38.25
1919	53 12		
1920	44 62		
1921	50.87		
1922	53.57		
1923	64 45		
1924	70 01		
1925	79 96		
1926	89.45		
1927	97.18	859 08 + 9 =	95.45
1928	105.92		
1929	119 04		
1930	119 62		
1931	113.45		

Advantages

1. The method is simple.
2. The result is entirely objective, i.e., it is not dependent upon individual estimate

Disadvantages

1. The semi-average method makes use of the arithmetic mean, which as we have seen before is greatly affected by extreme values. For this reason the semi-average trend line may be pulled out of its true position by such unusual occurrences as strikes, etc.
2. The method is used primarily for the fitting of straight line trends.

3. **Moving Average Method.** In the moving average method the trend is described by smoothing out the fluctuations of the data by means of a moving average.

The moving average is a series of successive averages secured from a series of items by dropping the first item in each group averaged and including the next in the series—thus obtaining the next average. To obtain a three item moving average, in the illustration below, the first three numbers (3, 5 and 7) are added (the total is entered in column 2 next to the middle item of the group). The first number (3) is then replaced by the next number in the column of figures (in this case 8) and the process is continued until the entire series has been included. Each total is then divided by three and the result is placed in column 3.

(1) Values	(2) 3 Item Moving Total	(3) 3 Item Moving Average
3		
5	15	5 00
7	22	7 33
10	29	9.67
12	36	12 00
14	41	13.67
15	46	15.33
17		

The fluctuations caused by the business cycle in an economic time series may be removed or partially eliminated by including in the moving average a number of items (years) equal to the length of the cycle which is evident in the data. The cyclical fluctuations will thus be smoothed out and a better measure of trend obtained.

Table 14 illustrates the application of such a moving average to the consumption of cigarettes in the United States. This table demonstrates the procedure to be followed in fitting a moving average consisting of both an odd number of items and an even number of items. The method for the odd period moving average consists of obtaining successive totals (in this case 7 items) by consecutively dropping the first item and adding the next in the series. The total for the first seven items in the illustration (table 14) equals 239.84. This sum is then placed next to the middle item of the group (1917). The second figure in column 3 (273.85) is obtained by dropping the figure for 1914 and adding the

figure for 1921. This process is continued until all the figures in column 2 have been included. Each of the figures in column 3 is then divided by 7 to obtain the seven year moving average entered in column 4.

Table 14—Computation of Trend—Moving Average Method
Consumption of Cigarettes in the United States, 1914-1932

(1) Year	(2) Consumption (Billions of Cigarettes)	(3) Seven Year Moving Total	(4) Seven Year Moving Average (Col. 3 divided by 7)
1914	16.86		
1915	17.96		
1916	25.29		
1917	35.33	239.84	34.26
1918	46.66	273.85	39.12
1919	53.12	309.46	44.21
1920	41.62	348.62	49.80
1921	50.87	383.30	54.76
1922	53.57	416.60	59.51
1923	64.15	452.85	64.70
1924	70.01	505.49	72.21
1925	79.96	560.54	80.08
1926	89.45	626.01	89.43
1927	97.18	681.18	97.31
1928	105.92	724.62	103.52
1929	119.04	748.24	106.89
1930	119.62		
1931	113.45		
1932	103.58		

Source: United States Department of Commerce.

When an even number of items is included in the moving average (as 6 years in table 15 below) the center point of the group will be between two years. It is, therefore, necessary to adjust or shift (known technically as center) these averages so that they coincide with the years. Columns 3 and 4 (six year moving average and six year moving total) may be obtained by the methods outlined for the odd period average as explained above. To center the values a two-item moving average is taken of the even item moving average.

A two year moving average is taken of the six year average. The resulting average is located between the two six year moving average values and therefore coincides with the years. The end result (a two year moving average of a six year moving average) is known as the six year moving average *centered*.

Advantages

1. Only simple computations are involved.
2. It may replace the fitting of complex mathematical curves.

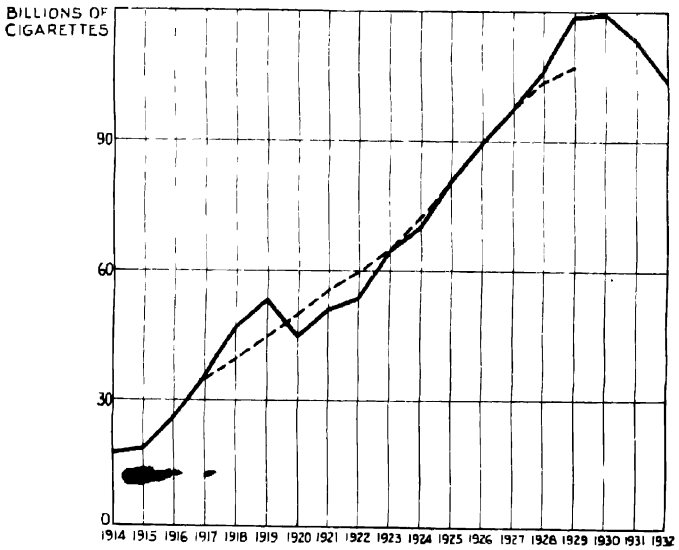
Disadvantages

1. It cannot be brought up to date. Depending upon the number of items included, the last point in the trend must

Table 15—Computation of Trend—Moving Average Method (Even Year Period)

Consumption of Cigarettes in the United States, 1914-1932

(1) Year	(2) Consumption (Billions of Cigarettes)	(3) Six-Year Moving Total	(4) Six-Year Moving Average	(5) Two-Year Moving Total of Col. 4	(6) Six-Year Moving Average Centered
1914	16.86				
1915	17.96				
1916	25.29	195 22	32 54	69.70	34.85
1917	35.33	222.98	37.16	79.81	39.91
1918	46.66	255 89	42.65	90 01	45.05
1919	53.12	284 17	47.36	99.58	49.79
1920	44.62	313.29	52.22	108.33	54.17
1921	50.87	336.64	56.11	116.69	58.35
1922	53.57	363.48	60 58	128.63	64.32
1923	64.45	408.31	68.05	143.82	71.91
1924	70.01	454.62	75.77	160.22	80.11
1925	79.96	506.97	84.50	178.09	89.05
1926	89.45	561.56	93.59	195.45	97.72
1927	97.18	611.17	101.86	209.30	104.65
1928	105.92	644.66	107.44	217.24	108.62
1929	119.04	658.79	109.80		
1930	119.62				
1931	113.45				
1932	103.58				



Source: United States Department of Commerce.

Fig. 15—Consumption of Cigarettes in the United States, 1914-1931. Trend indicated by a six year moving average.

occur several years before the end of the data. A five year moving average ends three years before the end of the data, a seven year average four years, etc.

2. The concept of trend involves the idea of a smooth growth or decline. The moving average is usually irregular in appearance.
3. A moving average fitted where the trend is that of a concave (upturning) curve will be higher than the true trend at all points, and lower in the case of a convex trend.
4. The moving average is computed by the use of the arithmetic mean. This form of average is greatly affected by extreme values. Because of this fact the moving average will be pulled decidedly out of line by such unusual events as strikes, disasters, etc.
5. The number of items giving the smoothest moving average is equal to the number of years included in the average length of the business cycle in the data. Since this average length must be estimated by the statistician the estimation will vary from person to person and, therefore, the method is not purely objective.

ADDITIONAL BIBLIOGRAPHY*

SUTCLIFFE, WILLIAM G., *Statistics for the Business Man*, pp. 196–200. Harper & Bros., New York, 1930.

* For readings in standard Statistics textbooks see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents

CHAPTER VI

TIME SERIES ANALYSIS—TREND
THE LEAST SQUARES METHOD—LINEAR

Formulae for Straight Lines

If a line is drawn on a graph its formula may be read by inspection.

The formula for line *d* in figure 16 is determined by obtaining the values of X and Y as indicated by the line itself.

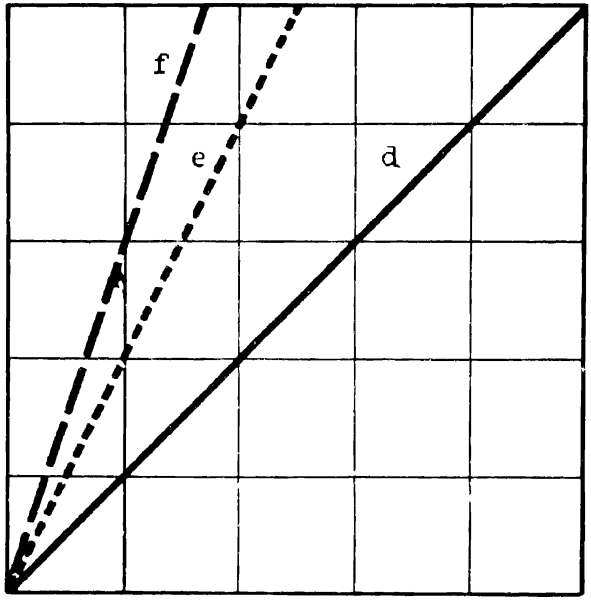


Fig. 16

For Line d

When X equals	Y equals
0	0
1	1
2	2
3	3

The formula for line *d* is, therefore,

$$Y = X.$$

The formula for lines *e* and *f* are obtained in exactly the same manner; viz:

Line *e* $Y = 2X$

Line *f* $Y = 3X$

The value of the coefficient of X (the number or constant which precedes X in the equation, as 3 in $Y = 3X$) indicates the number of units the value of Y will increase for each unit increase in X. For reference purposes this constant is given the letter b. The equation may now be written more generally as

$$Y = bX$$

The greater the value of the constant b the more rapid the rise in the line. The value of b therefore is the measure of the slope of the line. If the line of trend is downward, indicating a decrease in Y for each unit increase in X, the b value will be negative.

The Y Intercept

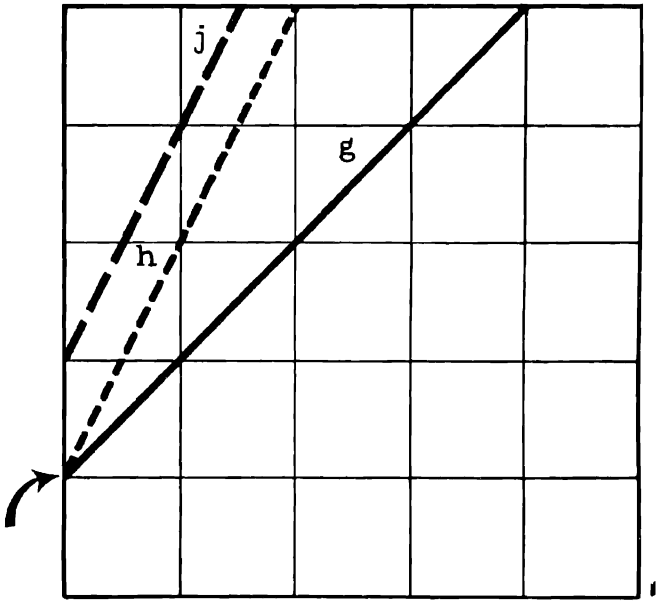


Fig. 17

Line g

If the values of X and Y for line g in figure 17 are obtained the result will be:

When X equals	Y equals
0	1
1	2
2	3
3	4
4	5

There is a unit increase in the value of Y for each unit increase in X. For this reason the value of b will be 1. It should be noted, however, that the Y value is constantly one unit greater than the X value, making the formula for this line of trend:

$$Y = 1 + 1X$$

The formula for line h is:

$$Y = 1 + 2X$$

since, in addition to Y being 1 when X is 0, it can be seen that for each unit increase in X there is a two unit increase in Y.

The value of the new constant is equal to the value indicated on the Y or vertical axis at the point at which the line crosses it, when X is equal to zero. This point is indicated by the arrow in figure 18 for lines g and h. The new constant 1 is assigned the letter a.

For Line j

When X equals	Y equals
0	2
1	4
2	6

The increase in Y again is two units for each unit increase in X. The slope constant b is therefore equal to 2. However, the line crosses the Y axis (when X equals zero) at 2 and as a result the a constant is equal to 2. The equation is then represented by:

$$Y = 2 + 2X$$

The constant a is sometimes called the **Y intercept** because the value is determined by the point at which the line crosses the Y axis (when X equals zero), while b indicates the **slope** of the line.

The formula for any straight line may be written generally as

$$Y = a + bX$$

The Least Squares Method

If a straight line trend is assumed, the line of trend will have a formula of the type:

$$Y = a + bX$$

In this formula the values of a and b must be determined. The formula $Y = a + bX$ will, however, describe any one of an infinite number of lines. It is necessary therefore to decide which line best describes the data. The principle of **least squares** aids in determining the line that *best* describes the trend of the data. The principle states that a line of best fit to a series of

values is a line the sum of the squares of the deviations (the differences between the line and the actual values) about which will be a minimum. There can be only one line having this qualification.

Least Squares Line

The least squares line for a given series may be obtained by the use of a set of "normal" equations. These "normal" equations are derived mathematically (see technical appendix IV) but for working purposes they may be obtained by multiplying the "type" equation, in this case[1]

$$Y = a + bX$$

through by the coefficients of each unknown (a and b). The coefficient of the first unknown (a) is 1. Multiplying the type equation through by 1 we have:

$$Y = a + bX$$

The formula must be summed up for all points. The summation results in

$$\text{(I)} \quad \Sigma(Y) = \Sigma a + b\Sigma(X)$$

But, the sum of a equals the number times the constant; viz,

$$\Sigma a = Na$$

since the sum of a number of constants will equal the constant multiplied by the number of times (N) it appears. The result may be written as:

$$\text{(I)} \quad \Sigma(Y) = Na + b\Sigma(X)$$

The coefficient of the second unknown (b) is X. Multiplying the type equation ($Y = a + bX$) through by X we obtain:

$$XY = aX + bX^2$$

This sums up to:

$$\text{(II)} \quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2)$$

By the use of these two equations the values of the two unknowns may now be determined and the trend line fitted.

Application of the Least Squares Method

The application of the least squares method for the determination of the trend for the production of aluminum in the United States is shown in table 16. The equation, since it is assumed to be linear (straight line), must be of the type

$$Y = a + bX$$

from which the two "normal" equations are obtained.

$$\text{(I)} \quad \Sigma(Y) = Na + b\Sigma(X)$$

$$\text{(II)} \quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2)$$

In order to solve for a and b the following values are necessary:

$$\Sigma(X);\ \Sigma(Y);\ \Sigma(XY);\ \Sigma(X^2);\ N$$

[1] The method outlined here is not a derivation but rather a short method for obtaining the necessary "normal" equations

Time is invariably placed on the X (horizontal) axis and for this reason constitutes the X variable, with production as the Y (vertical axis) variable.

The sum of X or $\Sigma\ (X)$ is obtained by adding the *numbers* of each year; for $\Sigma\ (Y)$ the production figures for all years are totaled.

The numbers of the years (as 1919, 1920, etc.) are inconvenient for calculating purposes and, since the numbering system for the years is merely arbitrary, the usual numbers are replaced with a simpler series of consecutive numbers. The correspondence between the two series must be noted by indicating the original number of the year to which the number of zero is now assigned. The year to which the number 0 is arbitrarily assigned is known as the origin year. The new numbering system is shown in column 2 of the work sheet below.

Table 16—Computation of Trend—Least-Squares Method
Annual Production of Aluminum in the United States, 1916-1930

(1) Year	(2) X	(3) Production of Aluminum (Millions of Pounds) Y	(4) XY	(5) X^2
1916	0	110 2	0	0
1917	1	143 3	143 3	1
1918	2	143.3	286 6	4
1919	3	134 5	403.5	9
1920	4	138 0	552 0	16
1921	5	55 0	275.0	25
1922	6	74 0	444 0	36
1923	7	129 0	903 0	49
1924	8	150 0	1200.0	64
1925	9	140.0	1260 0	81
1926	10	145 0	1450.0	100
1927	11	160.0	1760.0	121
1928	12	210 0	2520.0	144
1929	13	225.0	2925 0	169
1930	14	229 0	3206.0	196
	$\Sigma(X) = 105$	$\Sigma(Y) = 2186.3$	$\Sigma(XY) = 17328.4$	$\Sigma\ (X^2) = 1015$

N = the number of years (15 in illustrated problem) or items as the case might be.

Substituting the values obtained in the two normal equations:

$$\text{I}\quad \Sigma(Y) = Na + b\Sigma(X)$$
$$\text{II}\quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2)$$
$$\text{I}\quad 2186.3 = 15a + 105b$$
$$\text{II}\quad 17328.4 = 105a + 1015b$$

The equations may be solved simultaneously by obtaining equal values for the coefficient of one unknown (a or b).

If equation (I) is multiplied by 7:

$$\text{I} \quad 15304.1 = 105a + 735b$$
$$\text{II} \quad 17328.4 = 105a + 1015b$$

Equation one subtracted from equation two gives

$$2024.3 = 280b$$
$$\therefore b = \frac{2024.3}{280} = 7.23$$

Substituting 7.23 for b in the original equation (I) gives·

$$2186\ 3 = 15a + 105\ (7\ 23)$$
$$2186.3 = 15a + 759.15$$
$$1427.15 = 15a$$
$$a = 95\ 14$$

Having obtained the values for a and b, the normal equation can now be written with numerical coefficients and the formula for the line of trend written as:

$$Y = 95.14 + 7.23X$$

In interpreting the equation it is necessary to state the origin year and the units used in the enumeration of the original values.

The equation as finally stated will then read:

Trend of Annual Production of Aluminum
in the United States 1916–1930

$$Y = 95.14 + 7.23X$$

Year of origin 1916

Unit: in millions of pounds

Graphic Presentation of Trends

To obtain the various trend values of Y (in order that the trend line may be drawn on the graph) the various values of X indicated for each year on the work sheet are substituted in the equation. For 1918 the X value indicated in the work-sheet is 2. The X of the original equation is replaced by this value.

$$Y = 95.14 + (7.23)\ (2)$$
$$Y = 95.14 + 14.46$$
$$Y = 109.60 \text{ for the year 1918.}$$

This process is repeated until the trend value for each year is obtained. Since two points determine a straight line, in practice all that is needed to plot the line are two values for two different years.

Short Method for Computing Trend—Odd Number of Years

The procedure may be simplified when an odd number of years is used for the trend computation. The middle year may be taken as the origin year, thereby assigning to it an X value equal to zero. A *minus* sign is generally given to the X values for the years *previous* to the origin year and a *plus* sign to those *following* the origin year.

Applying this technique to the problem worked out above it will be found (see column 3 of table 17) that the sum of the X values will be zero, since they will consist of two like arithmetic progressions equal in amount but opposite in sign. This will modify the "normal" equations:

$$\text{I} \quad \Sigma(Y) = Na + b\Sigma(X)$$
$$\text{II} \quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2)$$

since $\Sigma(X) = 0$

$$a\Sigma(X) = 0$$
$$\text{and } b\Sigma(X) = 0$$

thus the normal equations are simplified to:

$$\text{I} \quad \Sigma(Y) = Na$$
$$\text{II} \quad \Sigma(XY) = b\Sigma(X^2)$$

and the need for a simultaneous solution is now eliminated.

Table 17—Computation of Trend—Least-Squares Straight Line—Odd Number of Years—Short Method

Annual Production of Aluminum in the United States, 1916-1930

(1) Year	(2) X	(3) Production of Aluminum (Millions of Pounds) Y	(4) XY	(5) X^2
1916	− 7	110 2	− 771.4	49
1917	− 6	143.3	− 859.8	36
1918	− 5	143 3	− 716.5	25
1919	− 4	134.5	− 538.0	16
1920	− 3	138.0	− 414 0	9
1921	− 2	55.0	− 110 0	4
1922	− 1	74.0	− 74.0	1
1923	0	129.0	.0	0
1924	1	150 0	150.0	1
1925	2	140.0	280.0	4
1926	3	145.0	435.0	9
1927	4	160 0	640.0	16
1928	5	210.0	1050.0	25
1929	6	225.0	1350.0	36
1930	7	229.0	1603.0	49
	$\Sigma(X) = 0$	$\Sigma(Y) = 2186.3$	$\Sigma(XY) = 2024.3$	$\Sigma(X^2) = 280$

Source: United States Bureau of Mines.

Substituting these values in the two simplified "normal" equations:

$$\text{I} \quad \Sigma(Y) = Na$$
$$\text{II} \quad \Sigma(XY) = b\Sigma(X^2)$$

the result is:

$$\text{I} \quad 2186.3 = 15a$$
$$\therefore \ a = 145.75$$
$$\text{II} \quad 2024.3 = 280b$$
$$\therefore \ b = 7.23$$

The resulting equation is written as follows:

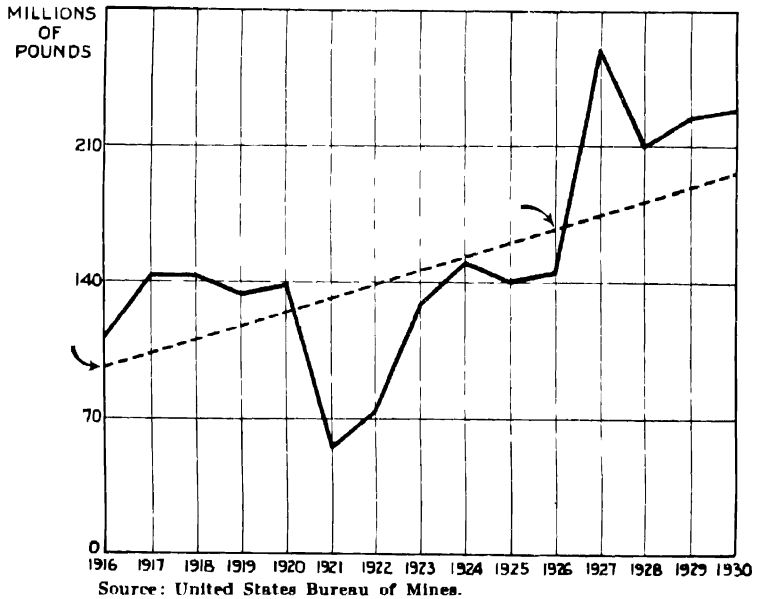
Trend of Annual Production of Aluminum
in the United States 1916–1930

$$Y = 145.75 + 7.23X$$

Origin: 1923

Unit: in millions of pounds.

Since a represents the value of Y when X equals zero, when a different zero point is used for X the a value is accordingly changed. As an example, the difference in the a value in the two trend equations (i.e, those obtained respectively by the long and short method) is due to the different points of origin. In the short method 1923 was taken as the origin year, in the long method



Source: United States Bureau of Mines.

Fig. 18—Annual Production of Aluminum in the United States 1916–1930. Trend indicated by a "least squares" straight line.

1916. The value of X was zero in 1923 for the short method, and zero in 1916 in the long method.

The equations can be shown to give the same results by using the corresponding X values for any year in both formulae.

If the production trend figure for 1923 is desired let $X = 7$. From table 16 substitute in the formula obtained the X value indicated for 1923.

$$Y = 95.14 + 7.23X \text{ Origin 1916}$$
(in millions of pounds)

with the result (for 1923):

$$Y = 95.14 + 7.23\ (7)$$
$$Y = 145.75$$

From table 17 for 1923 let $X = 0$ and substitute it in the formula as obtained by the simplified method:

$$Y = 145.75 + 7.23 \quad \text{Origin 1923}$$
(in millions of pounds)

and the result is (for 1923):

$$Y = 145.. \quad + 7.23\ (0)$$
$$Y = 145.75$$

Thus the two trend values correspond.

Shifting the Origin

A change in origin is merely a change in the starting point for the computation of the trend figure.

The b value (the measure of slope) is not affected by a shift in origin because no matter where the starting point be taken the line of trend will always have the same slope. In both equations in the previous paragraph the b (value) was 7.23 in spite of the differing origins (1916 and 1923).

The a value indicates the value of Y when X equals zero. The following method may be used to change the point of origin of the first equation (1916 origin) to a 1923 origin:

$$Y = 95.14 + 7.23X$$
origin 1916
(in millions of pounds)

In the new equation the origin is relocated at a point 7 years beyond the prior origin. By substituting 7 for X the value for Y must be equal to 145.75.

Since the slope is unaffected by a change in origin $b = 7.23$.

Substituting in the normal equation:

$$Y = a + bX$$
$$145.75 = a + 7.23X$$

But since $X = 0$ in 1923

$$\therefore \ a = 145.75$$

Summary of Steps for Shifting Origin

1: Use original equation	Y	$95.14 + 7.23X$ (origin 1916)
2: Substitute value for X for new origin year	X	7 since the 1923 trend value is desired
3: Obtain Y (trend) value	Y	145.75
4: Replace old a value with this new amount	Y	$145.75 + 7.23X$ (origin 1923)

Short Method—Even Number of Years

In applying the short method to a time series with an even number of years the fact that there is no middle year presents a difficulty. However, the middle point of the series may be used by assigning an origin between the two center years (as January 1, 1924).[1] The first year on either side of the origin will now be one half year away from the point of origin and as a consequence will be assigned as numbers + .5 or − .5, as the case may be. The second year on either side will be numbered + 1.5 and − 1.5. This numbering system is shown in table 18, column 3.

Since decimals or fractions are cumbersome the trend equation may be obtained more readily by working in terms of half years, thus eliminating decimal values (see column 4).

Table 18—Computation of Trend—Least Squares Straight Line Short Method —Even Year Period
Trend Expansion of F. W. Woolworth Co., 1916-1931

(1) Years	(2) Y Average Number of Stores	(3) X (In Years)	(4) X' (In Half Years)	$X'Y$	X'^2
1916	920	− 7.5	− 15	− 13800	225
1917	1000	− 6.5	− 13	− 13000	169
1918	1039	− 5.5	− 11	− 11429	121
1919	1081	− 4.5	− 9	− 9729	81
1920	1111	− 3.5	− 7	− 7777	49
1921	1137	− 2.5	− 5	− 5685	25
1922	1176	− 1.5	− 3	− 3528	9
1923	1261	− .5	− 1	− 1261	1
1924	1364	.5	1	1364	1
1925	1420	1.5	3	4260	9
1926	1484	2 5	5	7420	25
1927	1588	3.5	7	11116	49
1928	1727	4.5	9	15543	81
1929	1828	5.5	11	20108	121
1930	1890	6.5	13	24570	169
1931	1896	7.5	15	28460	225
	$\Sigma Y = 21922$	$\Sigma X = 0$	0	$\Sigma X'Y = 46612$	$\Sigma X'^2 = 1360$

Source: United States Department of Commerce; *Survey of Current Business.*

[1] When a figure for a year is indicated in a time series it represents the figure as of the middle of the year, or July 1 of that year.

These values may now be substituted in the two *simplified* normal equations:

$$\text{(I)} \quad \Sigma(Y) = Na$$
$$\text{(II)} \quad \Sigma(X'Y) = b\Sigma(X'^2)$$

Resulting in.

$$\text{(I)} \quad 21922 = 16a$$
$$a = 1370.13$$

The equation then reads:

Trend of Number of Stores in Woolworth Chain 1916–1931

$$Y = 1370.13 + 34.27X'$$

Origin: January 1, 1924
Unit: Number of stores
X' in ½ years

The equation in its present form is difficult to handle. For this reason it should be converted to the standard form which has as its origin July 1 of the year.

To shift the origin to the center of 1924 (July 1) it is necessary to obtain the trend value (Y) at the new point of origin and use that value for a.

The trend value for ½ year later may be obtained by substituting + 1 for X' in the equation (since X' is in terms of half years).

$$Y = 1370.13 + 34.27 \times 1$$
$$Y = 1404.40$$

Y may now be used as the new a value and the equation written·

$$Y = 1404\ 40 + 34\ 27\ X'$$

Origin 1924 (July)
X' in ½ years

Finally, an adjustment must be made to convert X' (in half years) to X (in years).

The symbol b used as the coefficient of X represents the increase per half year (in this case 34.27 stores). To obtain the increase per year b is doubled (68.54). The equation is now written:

Trend of Number of Stores in Woolworth Chain, (1916–1931)

$$Y = 1404.40 + 68\ 54\ X$$

Origin 1924
Unit: Number of stores.

Least Squares Method

Advantages

1. The method expresses trend in the form of a mathematical formula which may be easily interpreted.
2. Results obtained under the method are definite and independent of any subjective estimate on the part of the statistician.

3. The resulting equation is in convenient form for extrapolation (extension into future or past).

Disadvantages

1. The technique used is mathematical.
2. The method is based on the assumption that the data follows a trend that can be expressed by a mathematical equation.

ADDITIONAL BIBLIOGRAPHY*

FISHER, R. A., *Statistical Methods for Research Workers*, pp. 120–123. Oliver & Boyd, Edinburgh, 1932.

ODELL, C. W., *Educational Statistics*, pp. 189–199. Century Co., New York, 1924.

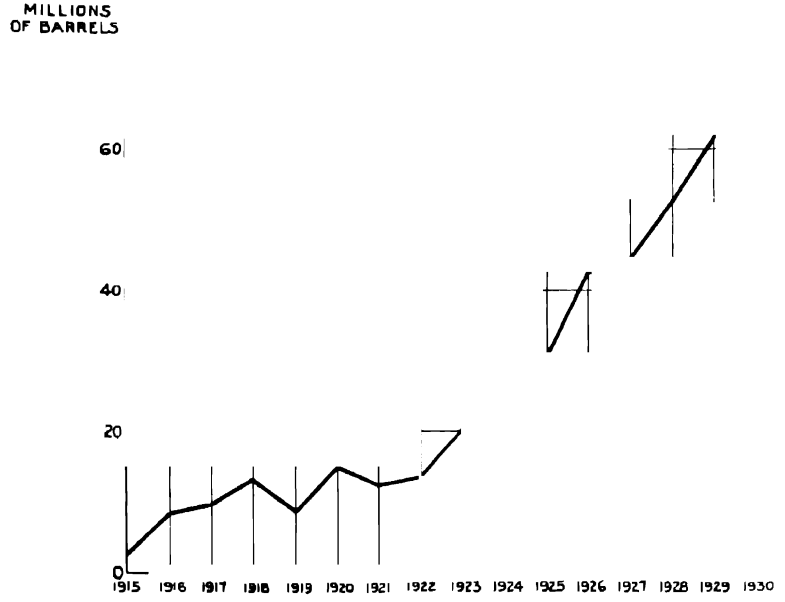
SUTCLIFFE, WILLIAM G., *Statistics for the Business Man*, pp. 201–216. Harper & Bros., New York, 1930.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER VII

TIME SERIES ANALYSIS—TREND
NON-LINEAR TRENDS

The straight line trend does not satisfactorily describe the trend of data which have a varying rate of growth. For example, in the illustration below only a curved line can accurately describe the trend of the data. The trend of gasoline exports is shown in figure 19.



Source: United States Bureau of Mines.

Fig. 19—Gasoline Exports from the United States, 1915-1930.

Methods of Fitting Non-Linear Trends:

a: **The Potential Series.** The parabola is the simplest type of curve used to describe the trend of data. The formula for a curve of the simplest parabolic type is:

$$Y = a + bX + cX^2$$

The fitting of a curve of this type follows closely the method of fitting a linear equation as explained in the previous chapter.

Here, however, there are *three* unknowns ("*a*", "*b*", "*c*") in the equations, necessitating three normal equations for its solution.[1]

Methods

1. Write down the "type" equation:

$$Y = a + bX + cX^2$$

2. Multiply each term of the equation by the coefficient of each unknown and sum up:

   a: For "Normal" equation one

   Multiply the type equation by the coefficient of "*a*" (which is 1) and sum up

   Equation (I) $\Sigma(Y) = Na + b\Sigma(X) + c\Sigma(X^2)$

   b: For normal equation two

   Multiply the type equation by the coefficient of b (which is X) and sum up

   Equation (II) $\Sigma(XY) = a\Sigma(X) + b\Sigma(X^2) + c\Sigma(X^3)$

   c: For normal equation three

   Multiply the type equation by the coefficient of "*c*", (X^2), and sum up

   Equation (III) $\Sigma(X^2Y) = a\Sigma(X^2) + b\Sigma(X^3) + c\Sigma(X^4)$

Table 19—Computation of Trend—Least Squares Method—Second Degree Parabola

Gasoline Exports from the United States, 1915-1930

(1) Years	(2) X	(3) Exports (Millions of Barrels) Y	(4) XY	(5) X^2	(6) X^2Y	(7) X^3	(8) X^4
1915	0	2.7	0	0	0	0	0
1916	1	8.5	8.5	1	8 5	1	1
1917	2	9.9	19.8	4	39 6	8	16
1918	3	13.3	39.9	9	119.7	27	81
1919	4	8.9	35.6	16	142.4	64	256
1920	5	15.3	76.5	25	382.5	125	625
1921	6	12.7	76.2	36	457.2	216	1296
1922	7	13.8	96.6	49	676.2	343	2401
1923	8	20.1	160.8	64	1286.4	512	4096
1924	9	28.3	254.7	81	2292.3	729	6561
1925	10	30.6	306.0	100	3060.0	1000	10000
1926	11	42.5	467.5	121	5142.5	1331	14641
192-	12	44.3	531.6	144	6379.2	1728	20736
1928	13	52.9	687.7	169	8940.1	2197	28561
1929	14	62.1	869.4	196	12171.6	2744	38416
1930	15	65.6	984.0	225	14760.0	3375	50625
	120	431.5	4614.8	1240	55858.2	14400	178312

Source: United States Bureau of Mines.

[1] In order to solve for a given number of unknowns as shown in an equation it is necessary to have the same number of equations involving these unknowns.

The resulting normal equations are:

Equation (I) $\Sigma(Y) = Na + b\Sigma X + c\Sigma(X^2)$

Equation (II) $\Sigma(XY) = a\Sigma(X) + b\Sigma(X^2) + c\Sigma(X^3)$

Equation (III) $\Sigma(X^2Y) = a\Sigma(X^2) + b\Sigma(X^3) + c\Sigma(X^4)$

From the values of X and Y

$\Sigma(X)$, $\Sigma(Y)$, $\Sigma(XY)$, $\Sigma(X^2)$, $\Sigma(Y^2)$, $\Sigma(X^2Y)$ and V can now be determined (as outlined in the previous chapter) and substituted in the normal equations shown above. A simultaneous solution is then used to obtain the desired values for the unknown constant.[1]

The fitting procedure is outlined in table 19.

(I) $\Sigma(Y) = Na + b\Sigma(X) + c\Sigma(X^2)$

$431.5 = 16a + 120b + 1240C$

(II) $\Sigma(XY) = a\Sigma(X) + b\Sigma(X^2) + c\Sigma(X^3)$

$4614\ 8 = 120a + 1240b + 14400c$

(III) $\Sigma(X^2Y) = a\Sigma(X^2) + b\Sigma(X^3) + c\Sigma(X^4)$

$55858\ 2 = 1240a + 14400b + 178312c$

Solving equations I and II to eliminate "a":

(II) $4614.8 = 120a + 1240b + 14400c$ (Equation II)

(I) $3236.25 = 120a + 900b + 9300c$ (Equation I times 7.5)

(IV) $1378.55 = 340b + 5100c$ Subtracting (1) − (2)

(III) $55858.2 = 1240a + 14400b + 178312c$ (Equation III)

(I) $33441.25 = 1240a + 9300b + 96100c$ (Equation I times 77.5)

(V) $22416.95 = 5100b + 82212c$

(IV) $20678.25 = 5100b + 76500c$ (Equation IV times 15)

$1738.70 = 5712c$

$c = .3044$

Substituting the value of c in equation IV

$1378.55 = 340b + 5100\ (.3044)$

$1378.55 = 340b + 1552.44$

$\therefore b = -1.5114$

Substituting the values for b and c in equation I

$4614.8 = 120a + 1240\ (-.5114) + 14400\ (.3044)$

$120a = 865.576$

$a = 7.2131$

[1] For review of the method of simultaneous solutions where there are more than two unknowns see any standard textbook on Elementary Algebra as Harding A M & Mullins G W, *College Algebra*

The final trend equation then reads:

Trend of Gasoline Exports from the United States, 1915–1930

$$Y = 7.2131 - .5114\ X + .3044\ X^2$$

Origin 1915

Unit: Millions of Barrels

The trend values for the various years may now be obtained by substituting the appropriate values of X (as indicated in column 2 table 19). Thus for 1925, substituting 10 for X

$$Y = 7.2131 - .5114\ (10) + .3044\ (10)^2$$
$$Y = 7.2131 - 5.114 + 30.44$$
$$Y = 32.5391 \text{ Millions of barrels}$$

The method can be further simplified when the data consists of an odd number of items. The middle year is selected as the year of origin. Therefore, $\Sigma(X)$ and $\Sigma(X^3)$ will then be equal to zero. The normal equations are:

$$\text{Equation (I)} \quad \Sigma(Y) = Na + b\Sigma(X) + c\Sigma(X^2)$$
$$\text{Equation (II)} \quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2) + c\Sigma(X^3)$$
$$\text{Equation (III)} \quad \Sigma(X^2Y) = a\Sigma(X^2) + b\Sigma(X^3) + c\Sigma(X^4)$$

However, since both $\Sigma(X)$ and $\Sigma(X^3) = 0$ the normal equations are reduced to.

$$\text{Equation (I)} \quad \Sigma(Y) = Na + c\Sigma(X^2)$$
$$\text{Equation (II)} \quad \Sigma(XY) = b\Sigma(X^2)$$
$$\text{Equation (III)} \quad \Sigma(X^2Y) = a\Sigma(X^2) + c\Sigma(X^4)$$

The values of a and c are then obtained simultaneously in the usual way, while b is obtained directly.

A more flexible curve than the second degree parabola may be obtained by using a parabola of a higher degree.[1]

The third degree parabola has the formula

$$Y = a + bX + cX^2 + dX^3$$

The general formula for this type of curve is

$$Y = a + bX + cX^2 + dX^3 + eX^4 \ . \ . \ . \text{ etc.}$$

These more elaborate forms of equations will generally tend to follow the data more closely but must be used with care if they are to describe the trend of the figures rather than the cyclical or seasonal movement.

The solutions for parabolas of a higher degree (dX^3, eX^4, etc.) may be arrived at in the same manner. The normal equations for the more complex formulas are obtained as before from the type equation and the values of the various unknown coefficients a, b, c, d, etc. secured through the method previously described

[1] The "degree" of an equation corresponds to the largest exponent in the equation

b: Exponential Series. Occasionally neither the straight line nor the parabola will be appropriate for describing the trend of a particular series. This occurs, for instance, where the trend is geometric in nature. One curve descriptive of the geometric type of trend has the formula:

$$Y = ab^x$$

This type of trend appears when the Y values tend to form a geometric progression [1] (such as the series 1, 2, 4, 8, 16, etc.) and the X values are arranged in the form of an arithmetic progression such as the series 2, 3, 4, etc. If plotted on semi-log paper (with logarithmic ruling on "Y" axis) a linear type of trend will make its appearance, and the resulting curve is therefore referred to as the **semi-logarithmic curve.** (See Chapter XVIII, p. 159).

If a geometric progression is formed by the "Y" values when the "X" values are arranged geometrically the formula is:

$$Y = aX^b$$

This type of curve will be linear on logarithmic paper (logarithmic rulings on both "X" and "Y" axis).

The formulae of the exponential type such as those above may be fitted readily by reducing them to logarithmic form.

Formula

$$Y = a\,b^x$$

reduced to logarithms reads

$$\log\ Y = \log a + X \log b$$

The *normal* equations may then be obtained as above.

There are a number of special exponential curves of some importance for trend purposes. One of the more important curves of this type is known as the Gompertz curve. The formula is:

$$Y = a\,b^{c^x}$$

ADDITIONAL BIBLIOGRAPHY*

FISHER, R. A., *Statistical Methods for Research Workers*, pp. 133–150. Oliver & Boyd, Edinburgh, 1932.

[1] A geometric progression is a series in which the values increase at a constant ratio

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER VIII

TIME SERIES ANALYSIS
SEASONAL AND CYCLICAL ANALYSIS

Seasonal Variation

Seasonal variation is the technical term given the more or less regular movements within the year recurring periodically year after year.

Each month has a typical position in relation to the rest of the year. The problem of seasonal variation is to determine this typical or average position of each month.

Methods of Measuring Seasonal Variation

The most generally used methods for measuring the seasonal variation occurring within a time series are the:

1. Simple average method.
2. Link Relative method.
3. Ratio to moving average method.
4. Ratio to trend method.

Table 20—Average Weekly Freight Car Loadings in U. S., 1919-1933

Year	Jan.	Feb.	March	Apr.	May	June	July	Aug	Sept	Oct.	Nov.	Dec.	Aver.
	(Unit 1000 Cars)												
1919....	728	687	697	715	759	809	858	892	960	967	807	758	803.8
1920..	820	776	848	731	862	860	901	968	969	1005	884	723	862.2
1921...	705	683	692	706	757	765	751	810	841	929	761	683	756.9
1922.. .	702	765	826	723	787	842	825	877	935	992	944	838	838.0
1923 ...	845	842	917	941	975	1011	986	1041	1037	1078	978	826	956.4
1924 ...	858	908	916	875	895	906	894	974	1037	1091	975	847	931.3
1925...	921	905	924	941	968	989	986	1080	1074	1107	1024	888	983.9
1926...	923	919	969	958	1037	1028	1049	1104	1148	1205	1068	904	1026.0
1927..	946	956	1002	975	1024	999	979	1062	1097	1115	956	834	995.4
1928....	862	897	951	935	1002	985	986	1058	1117	1175	1061	883	992.7
1929..	893	942	962	996	1051	1052	1038	1117	1135	1169	978	835	1014.0
1930	837	876	883	912	914	930	895	938	931	950	798	680	878.7
Total	10040	10156	10587	10408	11031	11176	11148	11921	12281	12783	11234	9699	
Averages	836.6	846.3	882.3	867.3	919.3	931.3	929.0	993.4	1023 4	1065.2	936.2	808.3	

Table 21 — Computation of Index of Seasonal Variation
Simple Average Method

Freight Car Loadings in the United States, 1919-1933

(1) Month	(2) Average for Month (from Table)	(3) Trend Correction	(4) Corrected Average	(5) Index of Seasonal
January	836.6	—	836 6	.92
February	846 3	- 1 1	844 9	93
March	882 3	2 8	879 5	96
April	867 3	- 4 2	863 1	.95
May	919 3	- 5 6	913 7	1 00
June	931 3	- 7 0	924 3	1 01
July	929 0	- 8 4	920 6	1 01
August	993 4	- 9 8	983 6	1 08
September	1023 4	- 11 2	1012 2	1 11
October	1065 2	- 12 6	1052 6	1 15
November	936 2	- 14 0	922 2	1 01
December	808 3	- 15 4	792 9	87
Total			10946 2	
Average			912 2	

The Simple Average Method

1. Average (arithmetic mean) the values for each month for all the years (see table 20). The result is the typical value for each of the twelve months.

2. Adjust for trend. Each of the averages just computed will be distorted by the secular trend of the data. If the trend is upward, December will be higher than it should be in relation to trend since it occurs later along the trend line.

The increase per month due to trend may be determined by fitting a "least squares line" to the *average monthly figures for each year* and dividing the *b* value (slope) by 12. The resulting value will then represent the amount each monthly average is distorted due to trend as compared to the previous month.

Thus to reduce the February average to the level of the first month, January, the amount of the trend increment may be subtracted from that average. To reduce March to the January level, it is necessary to subtract from it 2 times the trend increment, for April 3 times, etc. (see table 21, column 3).

3. The resulting corrected averages may then be expressed as a percentage of the average of the entire period (845 76). These values are known as the indices of seasonal variation. The figure of 93% for January means that the figure for January is typically 7% below the average for the year

Link Relative Method

The first step in the **link relative** method is to express the value for each month as a percentage of the previous month. From the data for Bituminous Coal Production in the United States 1914–1929 the figure for January, 1914 (40.19) is divided into the figure for February, 1914 (35.47), the figure for February, 1914 (35.47) is divided into that for March, 1914 (45.46) etc. The resulting percentage figures as 88.26% for February, 128.16% for March, etc., are called link relatives (see table 22).

Table 22—Link Relatives—Bituminous Coal Production in the United States, 1914-1929

Year	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1914		88.26	128.16	51.94	120.02	110.02	109.23	110.03	103.36	96.59	88.59	113.39
1915	99.29	78.00	108.46	94.25	103.24	109.76	104.74	107.28	107.34	107.91	101.22	102.39
1916	101.70	97.00	96.99	76.73	115.37	97.27	100.98	112.04	98.59	106.44	100.27	98.15
1917	108.78	86.20	115.77	87.42	112.52	99.43	98.87	102.33	95.23	107.16	98.66	92.35
1918	95.89	103.67	109.89	95.70	109.57	101.39	107.49	100.25	92.87	102.19	83.94	91.53
1919	105.00	76.08	106.82	95.39	116.75	98.69	115.23	100.41	110.55	118.65	33.23	195.90
1920	133.59	82.53	116.54	81.00	102.79	115.71	99.76	108.65	100.54	106.05	98.69	101.29
1921	77.27	76.60	98.51	90.66	120.99	101.70	89.64	113.66	101.64	124.59	82.37	85.98
1922	123.00	108.99	122.41	31.46	128.58	109.95	76.19	152.05	158.67	110.06	100.36	102.64
1923	108.01	84.04	111.00	90.91	107.26	98.72	99.21	108.29	94.58	106.42	87.27	92.82
1924	127.33	90.08	87.29	73.70	106.08	97.46	106.01	107.71	117.98	114.23	87.97	109.90
1925	111.61	75.08	96.52	89.55	105.28	104.76	106.19	113.39	104.32	113.64	95.45	104.00
1926	101.31	86.79	99.05	86.88	97.46	107.51	103.51	106.64	105.66	111.47	109.38	96.57
1927	99.09	93.01	113.68	57.65	102.37	102.88	92.11	123.96	100.53	104.96	92.33	101.66
1928	107.46	93.53	106.31	73.22	113.76	98.20	100.80	113.31	100.46	121.94	91.42	94.22
1929	120.19	91.87	83.24	93.75	108.91	94.77	106.74	108.01	101.42	115.10	89.16	101.12

The link relatives for each month (all the January's, etc.) are then averaged. The typical position of each month in relation to the previous month is thus obtained. For instance, June is typically 101.6% of May. Since the arithmetic mean of the monthly averages would be distorted by unusual monthly values, the median is used as the averaging method since it is less disturbed by extreme values.

The median (or typical) link relatives show the relation of each month to the month before but not to the rest of the months. Therefore, it is necessary to establish a relationship between these various links or convert the link relatives into a series of chain relatives. This is accomplished by arbitrarily setting the value of January as 100%. The median link for February is 87.5%, which indicates that February is typically that percent of January and therefore is 87.5% of 100%. The median link for March establishes its chain relative as 102.6% of the chain for February (87.5%) or 89.8%. This computation is continued by multiplying each of the median link relatives by the chain for the preceding month. This process is repeated for all of the twelve averages; and, in addition, for the second January median link which is the value indicated for the first January repeated (table 23, column 3).

Table 23—Computation of Seasonal Variation (Link Relative Method)

Month	(1) Median Link Relative	(2) Chain Relatives	(3) Adjusted Chain Relatives
January	107.5	100.0	100 0
February	87.5	87.5	86.9
March	102.6	89.8	88.7
April	87.2	78.3	76.6
May	109.2	85 5	83.3
June	101.6	86.9	84.1
July	102.3	88.9	85.6
August	108.5	96.5	92 6
September	101.5	97 9	93 4
October	109 0	106.7	101 7
November	91.9	98 1	92 5
December	101 2	99 3	93 2
January	107 5	106 7	100 0

A discrepancy exists between the chain relative for the first January and that for the second. The difference is due to the trend increment which makes each succeeding January higher or lower than that of the preceding year.

The difference between the two values (6.7) thus represents the trend increment. It is necessary to adjust the chain relatives for the effect of trend, therefore increasing multiples of one-twelfth of the discrepancy from each chain value—starting with 1/12 for February, 2/12 for March, etc., must be subtracted out. The chain relatives will be then reduced to the same level as January [1] (see table 23, column 3).

The end result is an index of seasonal variation with a base of January.

Ratio to Moving Average Method

1. The seasonal variations in the data are smoothed by means of a twelve month moving average. The differences between the actual values and this moving average are due to seasonal movements.

2. The ratio of each value to the corresponding moving average values for each month is then obtained.

3. The ratios are then averaged for each month of all the years, using either the mean or median for this purpose.

4. The resulting averages are the indices of seasonal variation.

Ratio to Trend Method

The ratio-to-trend method measures the seasonal variation and in addition the combined cyclical and residual variations.

[1] The trend discrepancy may be distributed on various other bases (see Mills, F. C., *Statistical Methods*, p 320).

Method

Express each actual monthly value as a percent of its corresponding trend value, as computed by the "least squares" method[1] (see table 24, column 4).

Since trend is 100% for all of these values, if plotted graphically the resulting graph would be that of the original data expressed in percentage form with trend removed.

The ratios $\left(\frac{\text{actual}}{\text{trend}}\right)$ are then averaged for each month over the entire period of years. If used in averaging the arithmetic mean may be distorted by extreme values, therefore these values are excluded before averaging.[2]

The extreme or unusual values may be located by means of a **multiple frequency table.** The multiple frequency table is a multi-column frequency distribution of the ratios (A/T) with one column for each month (illustrated in figure 20).

	JAN.	FEB.	MAR.	APR.	MAY	JUNE	JULY	AUG.	SEPT.	OCT.	NOV.	DEC.
2.50-2.74				/	ⓘ							
2.25-2.49				/	ⓘ							
2.00-2.24				/								
1.75-1.99			/	////	///	////	//	/				
1.50-1.74			/	/	////	/	/	///	/			
1.25-1.49			//	//	/	//	𝍸 /	///	//	/		
1.00-1.24	/		𝍸	///	///	𝍸 /	/	///	////	///		
.75- .99	/	////	//	/	/		//	//	////	////	////	𝍸 /
.50- .74	𝍸 /	𝍸 /	//			/	/	//	////	𝍸	𝍸 /	𝍸 //
.25- .49	𝍸	///	/				/	/		/	////	//
.00- .24	//	//	/	ⓘ	ⓘ	ⓘ	/			/	/	

Fig. 20—Multiple Frequency Table of Ratios of Building Contracts Awarded to Trend Values for Each Month, 1919--1933.

[1] The monthly trend values may be readily obtained by fitting the "least squares" line to the annual averages and dividing the "b" value by 12. Since the annual equation has its origin at the middle of the year (July 1) it will be necessary to shift the origin ½ month forward to center it on the month. This may be accomplished by adding ½ of the "b" value to "a".

[2] If the median is used for the averaging process it may not be typical when there are few items in the group averaged.

When the average is computed the unusual figures (indicated by circles in the multiple frequency table) are excluded from the average.

The resulting average ratios of actual values to trend will show the typical relation of each month to trend or the seasonal indices (see table 25). For example, the value of 102% for March means that that month is typically 2% above the trend for the month.

If the seasonal index for each month is subtracted from its respective ratio of actual to trend, the seasonal variation will be eliminated from the series, leaving as the only remaining fluctuations the combined cyclical and residual (random).

Table 24—Computation of Seasonal and Cyclical Variation—Ratio to Trend Method

Roadbuilding Contracts Awarded for Concrete Highways and Streets in the United States—1919-1933

(Two Years Shown Only)

(1) Year and Month	(2) Contracts Awarded (Million Square Yards)	(3) Trend	(4)	(5) Index of Seasonal	(6) Cyclical and Residual
	A	T	A/T	$1+S$	$C+R$
1919					
January	.27	5.17	.05	.51	− .46
February	.78	5.20	.15	.57	− .42
March	2.37	5.23	.45	1.02	− .57
April	5.01	5.26	.95	1.64	− .69
May	9.43	5.29	1.78	1.50	.28
June	6.61	5.33	1.24	1.37	− .13
July	5.75	5.36	1.07	1.18	− .11
August	8.15	5.39	1.51	1.16	.35
September	3.84	5.42	.71	.99	− .28
October	2.79	5.45	.51	.80	− .29
November	2.01	5.48	.37	.59	− .22
December	3.11	5.52	.56	.67	− .11
1920					
January	1.96	5.55	.35	.51	− .16
February	4.22	5.58	.76	.57	.19
March	6.25	5.61	1.11	1.02	.09
April	5.79	5.64	1.03	1.64	− .61
May	5.61	5.68	.99	1.50	− .51
June	2.94	5.71	.51	1.37	− .86
July	2.63	5.74	.46	1.18	− .72
August	2.04	5.77	.35	1.16	− .81
September	2.95	5.80	.51	.99	− .48
October	1.45	5.83	.25	.80	− .55
November	1.32	5.87	.22	.59	− .37
December	2.01	5.90	.34	.67	− .33

Source: Portland Cement Association.

Table 25—Computation of Seasonal Index—Ratio-to-Trend Method

Month	Monthly* Total (1919–1933)	Number of* Months Used	Monthly Average
January	7.65	15	.51
February	8.54	15	.57
March	15.40	15	1.02
April	23.02	14	1.64
May	18.11	12	1.50
June	19.29	14	1.37
July	17.78	15	1.18
August	17.39	15	1.16
September	14.86	15	0.99
October	12.08	15	.80
November	8.87	15	.59
December	10.04	15	.67

* Does not include "extreme" months, see multiple frequency table.

Ratio-to-Trend Method—Summary

1. Fit a trend line to the data.
2. Compute the ratio of each actual value to its respective trend (A/T).
3. Average the ratios for each month, using the arithmetic mean for averaging. First, however, eliminate extreme values located by means of the multiple frequency table. The resulting figures are the indices of seasonal variation.
4. Subtract the respective indices of seasonal variation from the ratios for each month. The resulting series represents the cyclical and residual fluctuations occurring in the series.

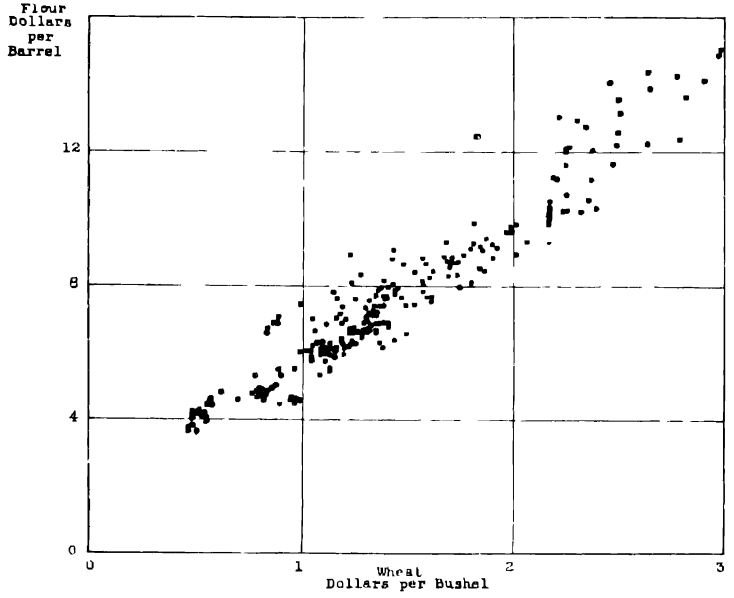
For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER IX

LINEAR CORRELATION

It is often desirable to observe and measure the relationship which occurs between two or more statistical series. It may, for instance, be desirable to know whether there is a relationship between changes in the cost of living and changes in wages; the grades on an examination and the intelligence quotient of a group of students; the amount of electrical current passed through a solution and the amount of substance deposited by electro-chemical reaction; the length of time elapsed and the amount of academic material retained in memory after various intervals of time; and many other similarly associated (correlated) series.

The relationship, or more accurately the association, between series may be established and measured by means of the correlation technique.



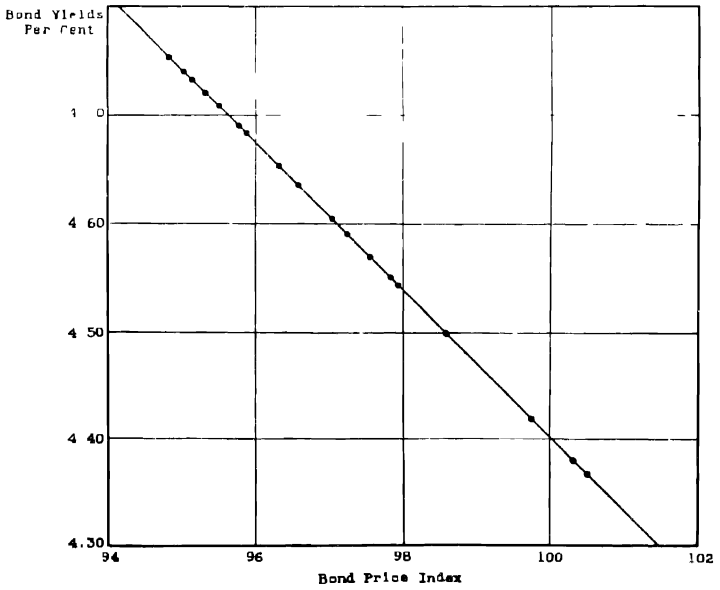
Source: Wheat prices—Daily Trade Bulletin.
Flour prices—Northwestern Miller.

Fig. 21—Scatter Diagram of Relation Between Wheat Prices and Flour Prices, by months, 1914—1933.

The Scatter Diagram

If two related (associated) series are plotted graphically with one variable placed on the X axis and the other on the Y axis, the result is known as the **scatter diagram.**[1] If there is a definite relationship resulting from plotting the associated variables on a chart the points will follow a definite line of movement or "path" as in figure 21.

If the relationship were perfect it is obvious that for every given value on the X axis, there would always be indicated a certain value on the Y axis. In a situation like this all the points would coincide with a curve or line instead of forming a path across the face of the scatter diagram. In figure 22 the figures of a bond *yield* index are plotted against those for the bond *price* index. Inasmuch as one series was computed from the other the relationship, of course, is perfect.



Source: Standard Trade and Securities Service, Standard Statistics Corporation.

Fig. 22—Scatter Diagram of Relationship Between Standard Statistics Index of Bond Yields and Index of Bond Prices.

When the series are imperfectly associated a definite value of Y will result when a given value of X is selected. In accordance

[1] The independent or causual variable is placed on the X axis while the dependent variable is placed on the Y axis

with the more or less imperfect relationship the variation will cause the points to depart from the indicated line or curve, creating a scatter. If there is a high degree of association the scatter will be confined to a narrow "path." The less perfect the relationship between the two sets of data, the greater will be the departures from the indicated line or course. These departures are known as a "scatter."

Line of Regression

The trend or direction of this movement may be defined by means of a "least squares" line or a curve (see Chapter VI).[1] The resulting curve is known as the **line of regression.** If the trend of the data is linear (non-linear regressions are treated in chapter X), the resulting equation will be of the type

$$Y = a + bX$$

The values of a and b are obtained from the "normal" equations:

$$\text{(I)} \quad \Sigma(Y) = Na + b\Sigma(X)$$
$$\text{(II)} \quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2)$$

as in the case of the straight line trend (see pp. 55).

Standard Error of Estimate

This equation is used to estimate a theoretical value of Y for a given value of X. If the relationship is not perfect the actual will not coincide with the theoretical values, because of the scatter or variation about the line. If the scatter is definitely measured the variation may then be allowed for and a range established within which all values will fall.

The measure used for this purpose, the **standard error of estimate,** is similar to the **standard deviation.** The standard deviation measures the variation or scatter about the **arithmetic mean,** while the standard error of estimate is a measure of the variation or scatter about the **line of regression.**

The standard deviation is the average (quadratic mean) of the deviations about the *arithmetic mean,* while the standard error of estimate is the average (quadratic mean) of the deviations about the *line of regression.*

$$S_y = \sqrt{\frac{\Sigma(d^2)}{N}}$$

Where

S_y = Standard error of estimate
d = deviation of actual values (Y) from theoretical (Y_c), or ($Y - Y_c$)

[1] Other methods of fitting such curves may be used, generally with less useful results (compare the freehand method)

The standard error of estimate may be used in the same manner as the standard deviation. One standard deviation measured off plus and minus about the arithmetic mean includes 68% of the cases; and one standard error of estimate will also include 68% of the cases when measured off plus and minus about the line of regression.[1]

Number of Standard Errors	Percent of Cases Included
$\pm$ .6745 S_y	50%
$\pm$ 1.0000	68%
$\pm$ 2 0000	95%
$\pm$ 3.0000	99.7%

Table 26—Computation of Coefficient of Correlation—Ungrouped Data
Circulation and Minimum Line Rates for National Advertising in 30 Daily Newspapers in New England, 1933

(1) Newspapers	(2) Circulation (Thousands) (X)	(3) Rate per Line (In Cents) (Y)	(4) (X Y)	(5) X^2	(6) Y^2	(7) Theoretical Regression Values Y_c	(8) $(Y - Y_c)$ d	(9) d^2
#1	166	33	5478	27556	1089	34	− 1	1
2	192	42	8064	36864	1764	38	+ 4	16
3	301	57	17157	90601	3249	55	+ 2	4
4	149	30	4470	22201	900	31	− 1	1
5	111	25	2875	13225	625	25	+ 0	0
6	108	23	2484	11664	529	25	− 2	4
7	446	75	33450	198916	5625	78	− 3	9
8	381	65	24765	145161	4225	68	− 3	9
9	399	70	27930	159201	4900	71	− 1	1
10	158	32	5056	204964	1024	33	− 1	1
11	451	79	35629	203401	6241	79	+ 0	0
12	133	27	3591	17689	729	29	− 2	4
13	108	22	2376	11664	484	25	− 3	9
14	154	30	4620	23716	900	32	− 2	4
15	331	47	10857	53361	2209	44	+ 3	9
16	150	32	4800	22500	1024	31	+ 1	1
17	403	70	28210	162409	4900	71	− 1	1
18	149	32	4768	22201	1024	31	+ 1	1
19	343	65	22295	117649	4225	62	+ 3	9
20	247	50	12350	61009	2500	47	+ 3	9
21	117	25	2925	13689	625	26	− 1	1
22	231	47	10857	53361	2209	44	+ 3	9
23	217	43	9331	47089	1849	42	+ 1	1
24	196	42	8232	38416	1764	39	+ 3	9
25	166	33	5478	27556	1089	34	− 1	1
26	124	25	3100	15376	625	27	− 2	4
27	182	35	6370	33124	1225	36	− 1	1
28	166	33	5478	27556	1089	34	− 1	1
29	112	28	3136	12544	784	26	+ 2	4
30	177	35	6195	31329	1225	36	− 1	1
	6468	1252	322227	1725088	60650			125

Source: Editor and Publisher, *International Yearbook* for 1933.

$$\sigma_y = \sqrt{\frac{\Sigma(Y^2)}{N} - \left(\frac{\Sigma Y}{N}\right)^2} = \sqrt{\frac{60650}{30} - \left(\frac{1252}{30}\right)^2} = 16.74¢$$

[1] It is assumed that there is a normal or approximately normal distribution of the values about the line of regression. For a more complete discussion of these percentage figures see chapter XI

The procedure for obtaining the "least squares" line of regression and the standard error of estimate for the data in table 26 is outlined below.

$$\text{(I)} \quad \Sigma(Y) = Na + b\Sigma(X)$$
$$\text{(II)} \quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2)$$
$$\text{(I)} \quad 1252 = 30a + 6468b$$
$$\text{(II)} \quad 322227 = 6468a + 1{,}725{,}088\ 0b$$
$$\text{(I)} \quad 269931 = 6468a + 1{,}394{,}500.8b \text{ (Equation I times 215.6)}$$
$$\text{Subtracting} \quad 52296 = 330{,}587b$$
$$b = .1582$$

Substituting the value of b in Equation I

$$1252 = 30a + 6468(.1582)$$
$$30a = 228.7624$$
$$a = 7.6254$$

Line of Regression

$$Y_c = 7.6254 + .1582\,X$$

where

Y_c = Minimum rate per line in cents.
X = Circulation in thousands.

The standard error may now be computed by first obtaining the theoretical regression values from this equation (column 7, table 26) and obtaining the difference between these and the actual values (column 8, table 26).

For instance the regression value for paper #3 with a circulation (X) of 301 is obtained as follows:[1]

$$Y_c = 7.6254 + .1582(301)$$
$$Y_c = 55¢$$

The standard error of estimate may then be computed from:

$$S_y = \sqrt{\frac{\Sigma(d^2)}{N}} = \sqrt{\frac{125}{30}}$$
$$= \sqrt{4.17} = 2.04¢$$

On the basis of the equation obtained:

$$Y = 7.6254 + .1582\,X$$

and with a standard error of estimate of 2.04¢, a new newspaper that reaches a circulation of 400,000 can be expected to have a minimum linage rate for national advertising of between 65¢ to 77¢ (3 standard errors—99.7% chances out of 100). 95 papers out of 100 (2 standard errors of estimate) with this circulation would have a minimum rate between 67¢ and 75¢.

[1] The symbol Y_c is used to differentiate the theoretical regression value from the actual value.

The range of values is obtained by substituting the circulation value for X:

$$Y_c = 7.6254 + .1582(400)$$

resulting in a theoretical value of 71¢ for the estimated linage rate. If 3 standard errors of estimate are then measured about this value 99.7% of all papers can be expected to fall within this range, while if 2 standard errors of estimate are added to and subtracted from this value 95% of the newspapers would be included within this range.

Coefficient of Correlation

The size of the standard error of estimate is a measure of the degree of association between series. The larger the value of the standard error of estimate the greater the scatter about the line of regression and, of course, the poorer the relationship.

Standard errors cannot always be directly compared however, this because the standard error is expressed in terms of the original unit of the Y variables, and frequently the units are different as in, for instance, the association of the intelligence quotient and examination grades compared with the association of the intelligence quotient and the number of words misspelled in a spelling test.

Very often a more limited range is possible in one of two variables. For instance, if there are only 20 words on the spelling test and the number spelled correctly is taken as the score, the variation about the regression line is limited by this range as compared with the wider variation permitted by the limits of 0 to 100% on an arithmetic test. If the standard error of estimate is divided by the standard deviation (of the Y values) the resulting value will be in percentage form. Both measures are in the same units, the factor of dispersion is made a constant, and thus both difficulties arising from differing dispersions and units are overcome.

$$\frac{S_y}{\sigma_y}$$

When the relationship is perfect there will be no deviations from the line of regression. S_y will then be equal to zero and result in a value of zero for the ratio. If the relationship is poor the value of the standard error will be larger, the limit of its value being that of the standard deviation. The ratio will thus attain as its other limit 1, or 100%.

A perfect relationship is indicated by a ratio equal to zero and an imperfect relationship by a value of 100%. Since this inverts the usual manner of thinking in regard to such subjects a more readily understandable value can be obtained by subtracting the ratio from one:

$$1 - \frac{S_y}{\sigma_y}$$

The new measure results in a value of 1.00 for a perfect relationship and a value of zero for a wholly imperfect relationship. A similar measure is termed the **coefficient of correlation** and is used as the comparative measure of association. The formula for the coefficient of correlation is

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

The coefficient of correlation will have the same limits as the value outlined above; viz., zero to 100%. The value of r for the problem outlined in table 26 is

$$r = \sqrt{1 - \frac{(2.04)^2}{(16\ 74)^2}} = .9925$$

Although the relationship be good or even perfect it may be inverse; that is to say, an increase in the value of X results in a corresponding decrease in the value of Y. Under these circumstances the line of regression slopes downward. The value of b, the coefficient of slope (also called the coefficient of regression), in this equation is then negative.

The sign of the coefficient of slope (or regression) is attached to r to indicate whether it is positive or negative.

The problem of the measurement of association may be divided into three sections:

1. The determination of the form of relationship—the **line of regression.**
2. The measurement of variation about the established form of relationship—the **standard error of estimate.**
3. The reduction of measurement of association to a relative basis—the **coefficient of correlation.**

Product Moment Method—Ungrouped Data

By the use of algebraic manipulation (see technical appendix) a much less arduous method for computing r, S_y and the line of regression [1] may be evolved from the fundamental formula for the coefficient of correlation:

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

The formula for the simpler method is[2]

$$r = \frac{p}{\sigma_x \sigma_y}$$

[1] It is assumed that the regression line is linear.
[2] For the derivation of the formula see technical appendix V.

Where, for ungrouped data

$$r = \frac{\Sigma(XY)}{N} - \left(\frac{\Sigma(X)}{N}\right)\left(\frac{\Sigma(Y)}{N}\right)$$

$$\sigma_x = \sqrt{\frac{\Sigma(X^2)}{N} - \left(\frac{\Sigma X}{N}\right)^2}$$

$$\sigma_y = \sqrt{\frac{\Sigma(Y^2)}{N} - \left(\frac{\Sigma Y}{N}\right)^2}$$

S_y can be obtained from a transposition of the original formula for r from

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

where r and σ_y^2 are known from the above computation

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

$$S_y^2 = \sigma_y^2 (1 - r^2)$$

$$S_y = \sigma_y \sqrt{1 - r^2}$$

Where the equation has its origin at the point of averages the line of regression may be determined from the formula[1]

$$y = r \frac{\sigma_y}{\sigma_x} x$$

since y and x are in terms of deviation from their respective means. Because the line of regression must pass through the point of averages,[2] the a (y-intercept) value will be zero when that point is used as the origin. With the origin now at the point of averages the customary equation (here expressed in terms of deviations from the means) is resolved into

$$y = a + b x$$

Where $a = 0$

and $b = r \frac{\sigma_y}{\sigma_x}$

therefore $y = r \frac{\sigma_y}{\sigma_x} x$

The more usual form of equation with an origin at zero (or in terms of actual values) may be determined by a transformation. Since

$$y = Y - \overline{Y}$$

$$x = X - \overline{X}$$

$$Y - \overline{Y} = r \frac{\sigma_y}{\sigma_x} (X - \overline{X})$$

[1] For mathematical proof see technical appendix VI.

[2] For proof see technical appendix VI

For the problem above (see table 26):

$$p = \frac{\Sigma(XY)}{N} - \frac{\Sigma(X)}{N}\frac{\Sigma(Y)}{N} =$$

$$\frac{322,227}{30} - \left(\frac{6468}{30}\right)\left(\frac{1252}{30}\right) = 1743.92$$

$$\sigma_x = \sqrt{\frac{\Sigma(X^2)}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{1,725,088}{30} - \left(\frac{6468}{30}\right)^2} = 104.97$$

$$\sigma_y = \sqrt{\frac{\Sigma(Y^2)}{N} - \left(\frac{\Sigma Y}{N}\right)^2} = \sqrt{\frac{60,650}{30} - \left(\frac{1252}{30}\right)^2} = 16.74$$

$$r = \frac{p}{\sigma_x \sigma_y} = \frac{1743\ 92}{(104.97)(16\ 74)} = .9925$$

$$S_y = \sigma_y \sqrt{1 - r^2} = 16.74\ \sqrt{1 - (.9925)^2} = 2.04$$

$$y = r\frac{\sigma_y}{\sigma_x}x$$

$$y = .9925\frac{16\ 74}{104\ 97}x$$

$$y = .1582\,x$$

$$Y - \overline{Y} = b\,(X - X)$$

$$Y - 41.73 = .1582\,(X - 215.6)$$

$$Y = 7.63 + .1582\,X$$

Product Moment Method—Grouped Data

The Correlation Table

Where there are very many items in the series to be analyzed for association the procedure outlined above is unsatisfactory. The use of it invites error in proportion to the multiplicity of computations while the work of calculation is very great.

Large groups of data may be handled more efficiently by first grouping them into the form of a double frequency distribution or **correlation table.** The procedure is given in table 27 for the association between the prices of wheat and flour.

The items are located in the various boxes by reference to the class intervals in which their X and Y values fall. Thus an item with an X value of \$.80 and a Y value of \$.65 will be located in the box indicated by the X class interval \$.80 – .89 and the Y class interval \$.60 – .69.

The calculations may be more readily carried out if, instead of trying to use the midpoints of the class intervals as the actual values, an arbitrary origin is selected for both X and Y values.

Table 27—Correlation Table—Wheat and Flour Prices by Months, 1914-1933

X = Wheat Price per Bushel in Dollars

Class Interval	Mid Point	Deviation d	Frequency f	fd	fd^2	.40 - .59	.60 - .79	.80 - .99	1.00 - 1.19	1.20 - 1.39	1.40 - 1.59	1.60 - 1.79	1.80 - 1.99	2.00 - 2.19	2.20 - 2.39	2.40 - 2.59	2.60 - 2.79	2.80 - 2.99	Total	$f(d_x d_y)$
						5	7	9	11	13	15	17	19	21	23	25	27	29		
						0	1	2	3	4	5	6	7	8	9	10	11	12		
						20	6	25	37	52	24	15	15	13	18	6	5	4	240	
					fd	0	6	50	111	208	120	90	105	104	162	60	55	48	1119	
					fd^2	0	6	100	333	832	600	540	735	832	1458	600	605	576	7217	
15.00 - 15.99	15.5	12	1	12	144															144
14.00 - 14.99	14.5	11	5	55	605															616
13.00 - 13.99	13.5	10	5	50	500															520
12.00 - 12.99	12.5	9	10	90	810															864
11.00 - 11.99	11.5	8	5	40	320															360
10.00 - 10.99	10.5	7	14	98	686															840
9.00 - 9.99	9.5	6	17	102	612															721
8.00 - 8.99	8.5	5	28	140	700															790
7.00 - 7.99	7.5	4	46	184	736															764
6.00 - 6.99	6.5	3	54	162	486															597
5.00 - 5.99	5.5	2	16	32	64															86
4.00 - 4.99	4.5	1	34	34	34															33
3.00 - 3.99	3.5	0	5	0	0															0
Total			240	999	5697															6335

Y = Flour Price per Barrel in Dollars

The deviations from the arbitrary origin should be measured in terms of class intervals, as in the short methods for calculating the mean and the standard deviation.

The necessary values for the determination of r by the "product moment" method now may be readily secured.

$$r = \frac{p}{\sigma_x \sigma_y}$$

where

$$p = \frac{\Sigma[f(d_x d_y)]}{N} - \frac{\Sigma(fd_x)}{N} \cdot \frac{\Sigma(fd_y)}{N}$$

$$\sigma_x = \sqrt{\frac{\Sigma(fd_x^2)}{N} - \left(\frac{\Sigma fd_x}{N}\right)^2}$$

$$\sigma_y = \sqrt{\frac{\Sigma(fd_y^2)}{N} - \left(\frac{\Sigma fd_y}{N}\right)^2}$$

The sum of $fd_x d_y$ is determined.

1. By securing the cross product of the indicated value of d_x and d_y for each box (inserted in the lower left hand corner of the box).
2. By multiplying the values obtained by the frequency con-

tained within the box (result entered in upper right hand corner of box).

3. By adding up for all boxes in the table.

It is to be noted that all calculations for r are carried out in terms of class intervals, not actual values, without reducing X and Y to actual units for the computation. The application of this technique for the series in table 27 is shown below.

$$p = \frac{\Sigma[f(d_x d_y)]}{N} - \frac{\Sigma(fd_x)}{N} \cdot \frac{\Sigma(fd_y)}{N} = \frac{6335}{240} - \left(\frac{1119}{240}\right)\left(\frac{999}{240}\right)$$

$$= 26\ 3958 - (4.6625)\ (4\ 1625) = 6\ 9881$$

$$\sigma_x = \sqrt{\frac{\Sigma f(d_x^2)}{N} - \left(\frac{\Sigma fd_x}{N}\right)^2} = \sqrt{\frac{7217}{240} - \left(\frac{1119}{240}\right)^2}$$

$$= \sqrt{30\ 0708 - 21.7389} = 2\ 8865$$

$$\sigma_y = \sqrt{\frac{\Sigma f(d_y^2)}{N} - \left(\frac{\Sigma fd_y}{N}\right)^2} = \sqrt{\frac{5697}{240} - \left(\frac{999}{240}\right)^2}$$

$$= \sqrt{23\ 7375 - 17\ 3264} = 2\ 5320$$

$$r = \frac{p}{\sigma_x\ \sigma_y} = \frac{6\ 9881}{(2\ 8865)\ (2.5320)} = .9561$$

The standard error of estimate may now be computed from

$$S_y = \sigma_y\sqrt{1 - r^2}$$

by first converting σ_y to the original units of its series through multiplying by the size of the Y class interval (in this case, \$1.00).

$$S_y = \$2\ 5320\sqrt{1-(\ 9561)^2} = \$.7419$$

The line of regression now may be obtained, using x and y in original, not class interval, values:

$$y = r\frac{\sigma_y}{\sigma_x}x$$

$$y = 9561\left(\frac{2.5320}{.5773}\right) = 4.1935\ x$$

but since

$$y = Y - \overline{Y}$$

$$x = X - \overline{X}$$

$$Y - 7.6625 = 4.1935\ (X - 1.4325) \text{ and } Y = 1.6553 + 4.1935\ X$$

Coefficients of Determination and Alienation

When the dependent variable (Y) is causally related to the independent variable (X) and both series consist of simple

elements of equal variability, r^2 measures the variance[1] in Y that is explained[2] by X. The measure (r^2) is then termed the **coefficient of determination** (the phrase **index of determination** is used where the correlation is curvilinear).

Just as the **coefficient of correlation** is a relative measure of the *degree* of association between two series, the **coefficient of alienation** is a comparative measure of the *lack* of association.

$$\text{Coefficient of Alienation} = \sqrt{\frac{S_y^2}{\sigma_y^2}}$$

The square of the coefficient of alienation may be interpreted in a similar fashion to the square of the coefficient of correlation and is known as the **coefficient of non-determination.**

$$1 = r^2 + k^2$$

and

$$k = \sqrt{1 - r^2}$$

Correction for Number of Cases

When the number of cases is small, the coefficient of correlation must be adjusted for exaggeration of its value and the standard error must be adjusted for an underestimate.

The correction formulae are:

$$\bar{S}_y^2 = S_y^2 \frac{(N-1)}{(N-2)}$$

$$\bar{r}^2 = 1 - (1 - r^2) \frac{(N-1)}{(N-2)}$$

where $\bar{S}_y$ and $\bar{r}$ are the corrected values.[3]

Other Correlation Methods

Correlation from Ranks

The method of measuring correlation from the ranks or position of the various items has proven particularly useful in education and psychology.

In this method the data are numbered according to their position, as shown in table 28.

The following formula (Spearman's)[4] may then be applied.

$$\rho = 1 - \frac{6\Sigma(D^2)}{N(N^2 - 1)}$$

[1] Variance is the technical term for the square of the standard deviation.

[2] See Ezekiel, M., *Methods of Correlation Analysis*, page 375, note 4.

[3] For more complete analysis of the formulae for S_y and r see Ezekiel, M., *Methods of Correlation Analysis*, p. 121.

[4] See Kelley, T. L., *Statistical Method*, pp. 191–194, for derivation of this formula.

Table 28—Illustrating Rank Method of Measuring Association
Hypothetical Grades of Fifteen Students on Two Examinations

Student Number	Examination 1		Examination 2		Difference in Rank (D)	(D^2)
	Grade	Rank	Grade	Rank		
1	100%	1	90%	3	2	4
2	98	2	95	1	1	1
3	95	3	89	4	1	1
4	91	4	87	5	1	1
5	90	5	93	2	3	9
6	85	6	86	6	0	0
7	83	7	80	7	0	0
8	82	8	79	8	0	0
9	81	9	76	10	1	1
10	80	10	77	9	1	1
11	70	11	72	11	0	0
12	65	12	60	14	2	4
13	63	13	62	13	0	0
14	60	14	50	15	1	1
15	50	15	63	12	3	9
						32

where

ρ is the measure of correlation*

D is the difference between the two ranks given for each individual

N equals number of individuals.

For the problem above

$$\rho = 1 - \frac{6(32)}{15(225 - 1)} = .953$$

In case of ties in rank either one of two methods of assigning ranks may be used. In the **bracket method** all are assigned the same rank; but the next individual is given the rank that would have been assigned if the ties had received successive ranks.

In the **mid-rank method**, a rank equal to that of the middle of the tie is assigned to all items with identical values.

Student	Grade	Ranks: Bracket Method	Ranks: Mid Rank Method
A	100%	1	1
B	95	2	3
C	95	2	3
D	95	2	3
E	94	5	5
F	92	6	6.5
G	92	6	6.5
H	90	8	8

* *The symbol ρ is the Greek small letter rho.*

Assuming that the original data constitute a normal distribution, the following relationship may be established between r (the coefficient of correlation) and ρ.

$$r = 2 \sin \left(\frac{\pi}{6} \rho \right)$$

Spearman's Footrule

When only a rough approximation of the correlation based on ranks is desired, Spearman's "footrule" formula may be used:

$$R = 1 - \frac{6\Sigma G}{N^2 - 1}$$

where G represents the *positive* differences in rank.

Correlation and the Time Series

When two time series are correlated, if there is a similar upward trend the higher items toward the end of the first series will be associated with the higher items in the second. Due to the influence of trend there will be an exaggeration of the correlation. In either case the long-time relationship would tend to overshadow the short time movements upon which attention particularly centers.

The following methods may be used to overcome the difficulty:

1. The deviations from trend may be correlated.
2. The first differences (deviation of each item from previous item in series) may be correlated.
3. The two series may be adjusted for trend.

Types of Correlation

I. **Simple Correlation:**

Two variables; one independent and one dependent.

A. **Linear Correlation:** The change in one variable is at a constant ratio to change in the other.

1. **Direct:** An increase in one variable is accompanied by an increase in the other.
2. **Inverse:** An increase in one variable is accompanied by a decrease in the other variable.

B. **Non-Linear (Curvilinear) Correlation:** The change in one variable is at an increasing or decreasing ratio, not at a fixed ratio, to a change in the other variable.

II. **Multiple Correlation:**

There are more than two variables. One variable is dependent while the other variables are independent.

Multiple correlation may be:

A. **Linear:** Some variables may be associated directly and others inversely.

B. **Non-Linear**

C. **Joint:** The relationship between various independent and dependent variables change with any change in another independent variable.

III. **Partial Correlation:**

Partial Correlation measures the association between an independent and a dependent variable. It allows for the variation associated with specified other independent variables.

ADDITIONAL BIBLIOGRAPHY*

KELLEY, TRUMAN L., *Interpretation of Educational Measurements*, pp 163–171. World Book Co., Yonkers, New York, & Chicago, Illinois, 1927.

ODELL, C. W., *Educational Statistics*, pp. 147–179. Century Co., New York, 1925.

OTIS, ARTHUR S., *Statistical Method In Educational Measurements*, pp. 175–205. World Book Co., Yonkers, New York, & Chicago, Illinois, 1926

REITZ, H. L. (Editor), *Handbook of Mathematical Statistics*, pp, 120–129; 150–165. Houghton Mifflin Co., New York, 1926

RUCH, G. M., & STODDARD, GEORGE R., *Tests and Measurements in High School Instruction*, pp. 355–363. World Book Co., Yonkers, New York, & Chicago, Illinois, 1927.

TRAUBE, MARION R., *Measuring Results in Education*, pp. 389–411. American Book Co., New York, 1924.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents

CHAPTER X

CORRELATION—NON-LINEAR, MULTIPLE, PARTIAL

A straight line of regression does not always satisfactorily describe the association between variables. Frequently the relationship is too complex to be described by means of a simple straight line and therefore a curve must be used.

For instance, if the association between rainfall and crop yield is examined it is found that beyond a certain point a doubling of the amount of rainfall will not result in a doubling of the crop yield. On the contrary, an approximate point could be established beyond which the yield would *decrease* in like proportion up to complete extinction of the crop.

Types of Regression Curves

The types of curves which may be used are similar to those described in the chapter on trend (see chapter VII).

The two most important types of curves are:

1. Potential curves of the type

$$Y = a + bX + cX^2 + \quad . \quad . \quad . \text{ etc.}$$

2. Exponential curves of the type

$$Y = ab^x$$

In terms of logarithms exponential curves may be divided into

a. Logarithmic

$$\log\ Y = a + b \log\ X$$
$$\log\ Y = a + b \log\ X + c \log\ X^2$$

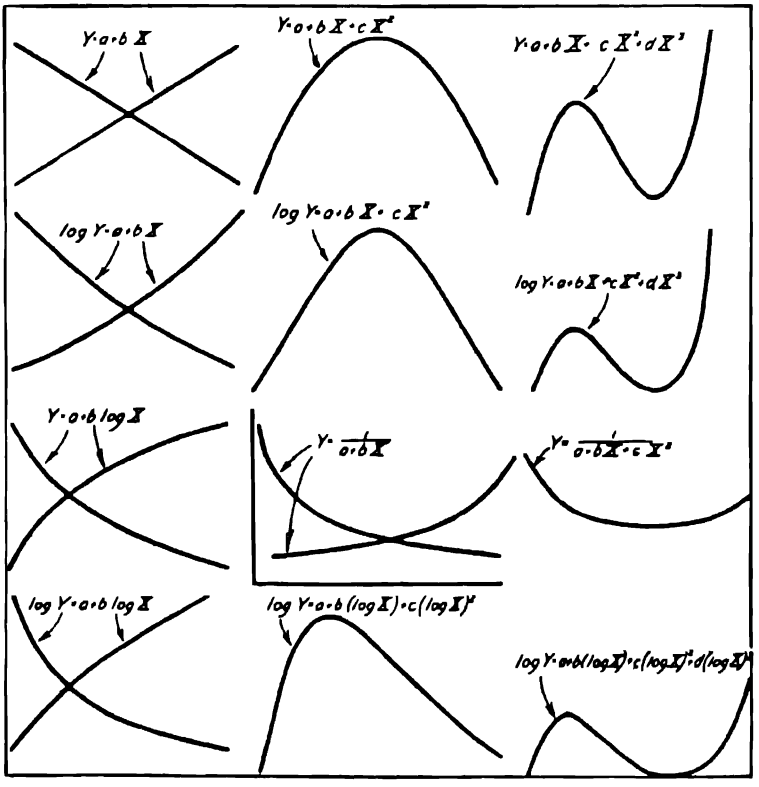
b. Semi-logarithmic

$$\log\ Y = a + bX$$
$$\log\ Y = a + bX + cX^2$$

or

$$Y = a + b \log\ X$$
$$Y = a + b \log\ X + c\ (\log\ X)^2$$

Figure 23 illustrates the appearance of some of these curves



From Ezekiel, Mordecai, *Methods of Correlation Analysis* p 69

Fig. 23—Curves Illustrating A Number of Different Types of Mathematical Functions.

The normal equations for any of the curves may be derived by the same method as outlined for trend curves (Chapter VI).

For a second degree parabola of the potential group

$$Y = a + bX + cX^2$$

The "normal" equations are

$$\text{(I)} \quad \Sigma(Y) = Na + b\Sigma(X) + c\Sigma(X^2)$$

$$\text{(II)} \quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2) + c\Sigma(X^3)$$

$$\text{(III)} \quad \Sigma(X^2Y) = a\Sigma(X^2) + b\Sigma(X^3) + c\Sigma(X^4)$$

For a semi-logarithmic equation of the type

$$\log\ Y = a + bX + cX^2$$

the normal equations are

$$\text{(I)} \quad \Sigma(\log Y) = Na + b\Sigma(X) + c\Sigma(X^2)$$
$$\text{(II)} \quad \Sigma(X \log Y) = a\Sigma(X) + b\Sigma(X^2) + c\Sigma(X^3)$$
$$\text{(III)} \quad \Sigma(X^2 \log Y) = a\Sigma X^2 + b\Sigma(X^3) + c\Sigma(X^4)$$

Before fitting an equation of the type

$$Y = \frac{1}{a + bX}$$

The equation is first converted to

$$\frac{1}{Y} = a + bX$$

or $$Y' = a + bX$$

where Y' is the reciprocal of Y.

The fitting of the curve may now be carried out in the usual fashion by obtaining the normal equations.

$$\text{(I)} \quad \Sigma(Y') = Na + b\Sigma(X)$$
$$\text{(II)} \quad \Sigma(XY') = a\Sigma(X) + b\Sigma(X^2)$$

Non-Linear Standard Error of Estimate

The standard error of estimate about the curve may be determined as before by obtaining the straight line (see page 76) from the formula:

$$S_y = \sqrt{\frac{\Sigma(d^2)}{N}}$$

or from [1]

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) - c\Sigma(X^2Y) - d\Sigma(X^3Y) - \text{etc.}}{N}$$

For types of curves other than the potential series this formula may be adapted by using the log of X or Y or their reciprocals as called for by the type equation. Thus for a curve of the type

$$\log Y = a + bX + cX^2$$

the formula would read

$$S_y^2 = \frac{\Sigma(\log Y)^2 - a\Sigma(\log Y) - b\Sigma(X \log Y) - c\Sigma(X^2 \log Y)}{N}$$

The Index of Correlation—rho (ρ)

When computed about a curve the comparative measure of correlation is known as the **index of correlation.** It is assigned the Greek letter "rho", or ρ, as a symbol.

$$\rho = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

The measure ρ is equal in value to r, the coefficient of correlation, when the regression is definitely linear. If the regression is non-

[1] The derivation of this formula follows along the same lines as that for the linear standard error outlined in technical appendix V.

linear, a curve will more closely approximate the data. As a result the deviations about the curve will tend to be smaller than those about the straight line. The standard error, therefore, will be smaller and will result in a larger value for rho. *Rho is always either equal to or larger than r.* Rho may be computed from the following formula:[1]

$$\rho^2 = \frac{a\Sigma(Y) + b\Sigma(XY) + c\Sigma(X^2Y) + \ldots - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

The **index of alienation** and the **indices of determination** and **non-determination** are computed in the same manner and with the same meaning as for the linear relationships (see page 84).

It is important to remember that the value for a given index of correlation (rho) may be compared to that for another association only when the same type of curve is used to describe both regressions.

At best non-linear regression lines must be used with extreme care, since the more complex the line the higher the index of correlation. An ultimate value of 100% may be reached where a curve so complex as to pass through all of the points is used. The resulting index would then be meaningless.

Correlation Ratio

The **correlation ratio** is only rarely used. In this measure the curve of regression is one which passes through the means of all the columns when the scatter diagram is divided into columns.

The regression is not defined by an equation.

Since the data must be divided into groups, the correlation ratio can be computed from the following formula:*

$$\eta = \sqrt{1 - \frac{\sigma^2_{ay}}{\sigma_y^2}}$$

where σ_{ay} is the standard deviation of the various values about the means of their respective columns.

The value of η is dependent upon the number of columns in the correlation table as compared to the number of cases used. With enough columns in the correlation table so that there is but one item in each column the correlation as computed by the correlation ratio would be perfect. A correction for fineness of grouping can be made by use of the formula:

$$\text{Corrected } \eta^2 = \frac{\eta^2 - \frac{(\kappa - 3)}{N}}{1 - \frac{(\kappa - 3)}{N}}$$

where

κ is the number of arrays or columns in the correlation table.

[1] See note [1] on page 91

* The symbol, η, is the Greek small letter eta

Since a curve passing through the means of the columns will be most descriptive of the data, the deviations about this curve will be smallest. The correlation ratio, therefore, will be either larger than or equal to the index of correlation or the coefficient of correlation. If the relationship is essentially linear the means will all coincide with a straight line and eta and r will be equal.

Since *rho* will also equal r if the regression is definitely linear:

$$\eta = \rho = r$$

but if it is non-linear:

$$\eta > \rho > r$$

Thus where the regression is basically non-linear, η would be larger than rho. A test for linearity of regression has been devised on this basis.

$$\zeta = \eta^2 - r^2$$

where ζ (zeta)* is the test for linearity.

When $\zeta = 0$, the regression is linear, when $\zeta \neq 0$ a non-linear regression is called for.

Method of Successive Elimination

Although in scientific experimentation it is possible to control all of the different variables and allow only the factor being studied to vary, this is not possible in many fields especially in the social sciences and business where numerous uncontrollable factors vary simultaneously. The relation of one of these numerous factors to a studied dependent variable to the exclusion of the other variables may be studied by means of the method of successive elimination.

The effectiveness of advertising as gauged by the number of returns secured to a keyed advertisement is dependent upon the circulation of the newspaper in which it was inserted and the size of the advertisement[1] among other factors. If it is desired to study the effectiveness of various sizes of advertisements, allowing for or more exactly excluding the effect of the differing circulation of the various papers in which the advertisements appeared, the relationship between circulation and returns is studied by means of either linear or non-linear correlation. The line of regression is determined by the usual methods. Since this variable is not the only factor determining the number of returns the points (actual values) will be scattered about the line of regression rather than coincide with it.

The corresponding theoretical values due to varying circulations may now be determined from the line of regression and the differences secured between the actual number of returns and these theoretical values. These values are termed the residuals (z').[2]

$$z' = Y - Y_c$$

* The symbol ζ is the Greek small letter zeta.

[1] Assuming that the advertising copy is the same for all advertisements.

[2] Some of these values will be negative showing below average returns for a paper of the specified circulation.

The residuals may now be correlated with the size of the advertisement to determine their relationship. The line of regression will express the variation in the returns if the circulation of the paper were held constant.

Another effective method is to use the line of regression to adjust the given data so that the returns are expressed as though all were obtained from a paper of a given circulation (as 100,000). An allowance is made for papers with smaller circulations by means of the regression line and an adjustment is made for larger circulations. The b value (in a linear regression) expresses the increase in returns per unit of circulation and can be used for this purpose.

Multiple Correlation

The fluctuations in a given series are seldom dependent upon a single factor or cause. The measurement of the association between such a series and several of the variables causing these fluctuations or associated with the dependent variable is known as **multiple correlation.**

Multiple correlation consists of the measurement of the relationship or association between a dependent variable and two or more independent variables. The procedure is similar to that for simple correlation with the exception that other variables are added to the regression equation. For two independent variables the regression equation, if linear, is of the type:

$$X_1 = a + b_{12 \cdot 3} X_2 + b_{13 \cdot 2} X_3$$

In the equation X_1 is the dependent or estimated variable (replacing the symbol Y previously used) and X_2 and X_3 are the independent variables.

The coefficient of regression or slope, $b_{12 \cdot 3}$, indicates the number of units change in the dependent variable for a given unit change in X_2, while $b_{13 \cdot 2}$ indicates the change in X_1 for a unit change in X_3. However, in the computation of these coefficients of regression the associations of each of the other independent variables with the dependent variable was taken into consideration. The coefficients, therefore, indicate the *net* relationship between the dependent and an independent variable, allowing for the other factors or variables which are also considered in the equation. The subscripts after the period indicate the other variables included. Thus $b_{12 \cdot 345}$ would give the net regression of variable X_2 in relationship to X_1, allowing for X_3, X_4, and X_5. The last named coefficients are therefore known as the **coefficients of net regression.**

The values for the coefficients may be obtained in the usual manner by making use of the previously outlined method (page 53) of obtaining the "normal" equations. For a linear relationship with three independent variables:

$$X_1 = a + b_{12 \cdot 34} X_2 + b_{13 \cdot 24} X_3 + b_{14 \cdot 23} X_4$$

the normal equations would be

$$\text{(I)} \quad \Sigma(X_1) = Na + b_{12\cdot34}\,\Sigma(X_2) + b_{13\cdot24}\,\Sigma(X_3) + b_{14\cdot23}\,\Sigma(X_4)$$

$$\text{(II)} \quad \Sigma(X_1X_2) = a\Sigma(X_2) + b_{12\cdot34}\,\Sigma(X_2^2) + b_{13\cdot24}\,\Sigma(X_2X_3) + b_{14\cdot23}\,\Sigma(X_2X_4)$$

$$\text{(III)} \quad \Sigma(X_1X_3) = a\Sigma(X_3) + b_{12\cdot34}\,\Sigma(X_2X_3) + b_{13\cdot24}\,\Sigma(X_3^2) + b_{14\cdot23}\,\Sigma(X_3X_4)$$

$$\text{(IV)} \quad \Sigma(X_1X_4) = a\Sigma(X_4) + b_{12\cdot34}\,\Sigma(X_2X_4) + b_{13\cdot24}\,\Sigma(X_3X_4) + b_{14\cdot23}\,\Sigma(X_4^2)$$

By assuming the origin of the equation to be at the point of averages this equation reduces to:[1]

$$\text{(I)} \quad p_{12} = b_{12\cdot34}\,\sigma^2_2 + b_{13\cdot24}\,p_{23} + b_{14\cdot23}\,p_{24}$$

$$\text{(II)} \quad p_{13} = b_{12\cdot34}\,p_{23} + b_{13\cdot24}\,\sigma^2_3 + b_{14\cdot23}\,p_{34}$$

$$\text{(III)} \quad p_{14} = b_{12\cdot34}\,p_{24} + b_{13\cdot24}\,p_{34} + b_{14\,23}\,\sigma^2_4$$

The equations may now be solved simultaneously to arrive at the desired values for $b_{12\cdot34}$, $b_{13\cdot24}$, and $b_{14\cdot23}$.

The standard error of estimate may now be computed from $S_{1\cdot234} = \sqrt{\frac{\Sigma d^2}{N}}$ or more readily from [2]

$$S^2_{1\cdot234} = \sigma^2_1 - b_{12\cdot34}\,p_{12} - b_{13\cdot24}\,p_{13} - b_{14\cdot23}\,p_{14}$$

and the coefficient of multiple correlation from

$$R_{1\cdot234} = \sqrt{1 - \frac{S^2_{1\cdot234}}{\sigma^2_1}}$$

or

$$R^2_{1\cdot} \qquad b_{12\cdot34}\,p_{12} + b_{13\cdot24}\,p_{13} + b_{14\cdot23}\,p_{14}$$

The coefficients of multiple alienation, determination, and non-determination may now be computed and applied as in simple correlation.

The same technique may be used for non-linear multiple correlation, using the general equation:

$$X_1 = a + f(X_2) + f(X_3) + f(X_4) + \text{etc.}$$

where "$f(X_2)$" indicates any function of X_2 such as a parabola, etc. As the complexity of the function increases the amount of computation required soon becomes so great as to render the use of the technique prohibitive.[3]

[1] See technical appendix VII—"a" may be obtained from the first "normal" equation

$$\Sigma(X) = Na + b_{12.34}\,\Sigma(X_2) + b_{13.24}\,\Sigma(X_3) + b_{14.23}\,\Sigma(X_4)$$

[2] See technical appendix VIII.

[3] A less arduous technique for curvilinear multiple correlation is outlined in Ezekiel, M., *Methods of Correlation Analysis*.

Partial Correlation

Coefficient of Partial Correlation

When it is desired to compute the separate or net effect or importance of each independent variable the technique of partial correlation may be used.

The coefficient of partial correlation is a relative measure of the association between the dependent variable and a given independent variable, eliminating the effect of the other independent variables.

There are a number of formulae which may be used to compute the association.

1. $r_{13\cdot24} = \sqrt{b_{13\cdot24} \cdot b_{31\cdot24}}$

2. $r_{12\cdot3} = \dfrac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}}$

3. $r_{14\cdot23} = 1 - \dfrac{(1 - R^2_{1\cdot234})}{(1 - R^2_{1\cdot23})}$

Coefficient of Part Correlation

A similar coefficient somewhat easier to compute is called the **coefficient of part correlation** ($_{12}r_{34}$).

The differences between these two coefficients may be shown by comparing the coefficient of partial correlation to that resulting from the correlation of:[1]

$(X_2 - b_{23\cdot4}\, X_3 - b_{24\cdot3}\, X_4)$ with $(X_1 - b_{13\cdot4}\, X_3 - b_{14\cdot3}\, X_4)$

while the coefficient of part correlation may be compared to the correlation between

$$X_2 \text{ and } (X_1 - b_{13\cdot24}\, X_3 - b_{14\cdot23}\, X_4)$$

The coefficient of part correlation may be computed from

$$_{12}r^2_{34} \qquad \frac{b^2_{12\cdot34}\, \sigma^2_2}{b^2_{12\cdot34}\, \sigma^2_2 + \sigma^2_1\, (1 - R^2_{1\cdot234})}$$

The subscripts to the right of the letter indicate the variables excluded.

Beta Coefficients

The relative importance of the individual independent variables in a multiple correlation in determining the dependent variable may be determined through resort to the *beta coefficients.*

The coefficients of regression of the multiple correlation regression equation indicate the increase in the dependent variable resulting from a unit increase in the indicated independent variable. However, the various independent variables are often expressed in different units making a direct comparison of the

[1] Suggested by Ezekiel in *Methods of Correlation Analysis*, Page 183

coefficients impossible. The coefficients of multiple regression may be made comparable by dividing each variable by its own standard deviation. Thus the original multiple regression equation

$$X_1 = a + b_{12\cdot3}X_2 + b_{13\cdot2}X_3$$

becomes

$$\frac{X_1}{\sigma_1} = a + b_{12\cdot3}\frac{X_2}{\sigma_2} + b_{13\cdot2}\frac{X_3}{\sigma_3}$$

and therefor

$$\beta_{12\cdot3} = b_{12\cdot3}\frac{\sigma_2}{\sigma_1}$$

$$\beta_{13\cdot2} = b_{13\cdot2}\frac{\sigma_3}{\sigma_1}$$

The beta coefficients are comparable measures and indicate the increase in the dependent variable resulting from an increase of *one standard deviation* in each independent variable.

ADDITIONAL BIBLIOGRAPHY*

BOWLEY, ARTHUR L., *Elements of Statistics*, pp. 350–408. P. S. King & Son, London, 1907.

FISHER, R. A., *Statistical Methods for Research Workers*, pp. 152–223. Oliver & Boyd, Edinburgh, 1932.

KELLEY, TRUMAN L., *Interpretation of Educational Measurements*, pp. 186. World Book Co., Yonkers, New York, & Chicago, Illinois, 1927.

ODELL, C. W., *Educational Statistics*, pp. 200–213; 245–279. Century Co., New York, 1925.

OTIS, ARTHUR S., *Statistical Method in Educational Measurements*, pp. 206–215. World Book Co., Yonkers, New York, & Chicago, Illinois, 1926.

RIETZ, H. L., (Editor), *Handbook of Mathematical Statistics*, pp. 129–149. Houghton Mifflin Co., New York, 1924.

SUTCLIFFE, WILLIAM G., *Statistics for The Business Man*, pp. 224–228. Harper & Bros., New York, 1930.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER XI

CORRELATION OF ATTRIBUTES

The previous section dealt with the measurement of the association between two measured characteristics of a given set of items. Two quantitative values were determined for each item and a coefficient of correlation was computed for the various paired values. Thus if the height and weight of a group of individuals are measured the coefficient of correlation can be computed. It will show the degree to which the greater heights are associated with the greater weights.

It is not always possible to use measurements for various characteristics. For instance, many classifications are qualitative such as light and dark, good and poor, etc. The association between heights and weights may be determined by classifying the individuals as light and heavy and tall and short. The association between the two characteristics may be determined by cross classifying each individual in a fourfold table of the type shown below, where a, b, c, and d represent the number of cases with each of the given pairs of characteristics.

	Short	Tall	Total
Light .	a	b	
Heavy .	c	d	
Total .			N

If the association is perfect—that is, if all tall people are heavy and all short people are light—all of the cases will be located in two boxes (cells)—in this case a and d. If there is absolutely no association—that is, if it is a matter of indifference insofar as weight is concerned if a person is tall or short—the cases will be distributed at random throughout the four boxes. Since there is an equal likelihood of a case appearing in any box there will tend to be an equal number in each box.

In a similar manner the qualitative characteristics of a group of individuals or items may be arranged in a table when more than two alternative attributes are dealt with. A table of this type is shown below:

	A	B	C	D	Total
A′ B′ C′ D′		n_{rc}			n_{r1} n_{r2} n_{r3} n_{r4}
Total	n_{c1}	n_{c2}	n_{c3}	n_{c4}	N

where $ABCD$ and A', B', C', D', are two sets of qualitative specifications.[1]

In this type of table if the association is perfect the cases will all fall in a diagonal row of cells if the table is symmetrical, for quality A will always accompany quality A', B will accompany quality B', etc. If there is no correlation there will be a tendency towards an equal distribution of cases among various boxes.

It is possible on the basis of these facts to compute a coefficient of association which will serve as a comparative measure of correlation.

Coefficient of Contingency

This coefficient of association (the coefficient of mean-square contingency) is based upon a comparison of the number of cases actually occurring in a given cell or box and the number of cases which would occur in the cell due to chance or a comparison of the actual distribution and the distribution occurring when there is no association.

If n_r is the number of cases in a given row, n_c is the number in a given column and n_{rc} is the number in a given cell

$$n_{rc} - \frac{n_r n_c}{N}$$

will be the difference between actual number of cases and the number of cases occurring due to chance. But the ratio of the square of this value to the theoretical number of cases is χ^2 (test for goodness of fit)[2] or

$$\chi^2 = \left[\frac{\left(n_{rc} - \frac{n_r n_c}{N} \right)^2}{\frac{n_r n_c}{N}} \right]$$

Pearson's **mean-square contingency,** ϕ^2 is obtained by dividing this value by N*

$$\phi^2 \quad \frac{\chi^2}{N}$$

[1] The qualitative specifications should be arranged in order where there are more than two such as poor, fair good, etc

[2] See page 109

* ϕ is the Greek letter phi.

and the coefficient of **mean-square-contingency** is

$$C\,C \quad \sqrt{\frac{\phi^2}{1+\phi^2}} = \sqrt{\frac{\chi^2}{N+\chi^2}}$$

A simpler form of this formula as shown by Yule is

$$C \quad \sqrt{\frac{S-N}{S}}$$

where

$$S = \Sigma\left(\frac{n^2_{rc}}{n_r n_c}\right)$$

or

$$S = N\,\Sigma\left(\frac{n^2_{rc}}{n_r n_c}\right)$$

Steps in Computation

1. Square the value in each cell (n_{rc}^2)
2. For each box multiply the number in that row by the number in that column (n_c).
3. Divide the squared value in each box $(n_{rc})^2$ by the corresponding $n_r n_c$.
4. Sum up for all rows and divide by N. The resulting value is S.
5. Substitute in[1]

$$C = \sqrt{\frac{S-N}{S}}$$

Fourfold Tables (2 x 2 classifications)

When two variates are classed in alternate categories, A and not A, B and not B.

1. Yule's coefficient of association and coefficient of colligation [2]

The coefficient of association is

$$Q = \frac{ad - bc}{ad + bc}$$

The coefficient of colligation is*

$$\omega = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

where a, b, c, d represent the frequencies contained within the various cells as follows:

a	b
c	d

[1] It has been shown that the number of classifications in the table will effect the largest possible value of C. When a table of 2 x 2 classifications is used C cannot exceed .707, when 4 x 4 the maximum value is .866 when 5 x 5 it cannot exceed .894, when 10 x 10 the maximum value is .949

[2] Objections have been offered to the use of both of these measures

* ω is the Greek letter Omega

When the relationship is perfect all of the cases will be concentrated in two of the boxes either ad or bc and therefore both Q and W will be equal to 1.00 or − 1.00. When there is no association the distribution of cases will be equal ($a = b = c = d$) and therefore both Q and W will equal zero.

2. *Pearson's cosine method.*

The coefficient of correlation by the cosine method for a four-fold table is

$$r = \cos \frac{\sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \pi$$

This coefficient will vary from $r = 0$ to $r = 1.00$. When the association is perfect there will be frequencies in only two of the squares (either a and d or b and c) and therefore $\frac{\sqrt{ad}}{\sqrt{ad} + \sqrt{bc}}$ will equal zero and r will equal 1.00. When there is an equal distribution of cases $a = b = c = d$ the fraction will equal .50 and $r = .00$.

Biserial Coefficient of Correlation

When the table is of $2 \times N$ classifications, i.e., when there are only two possible categories of one attribute and a number of classifications for the other attribute, the biserial coefficient of correlation may be computed. This type of classification is common since classification by sex and other similar two-fold categories appear frequently. It assumes that the distribution of the attribute is approximately normal.

The formula is

$$\text{biserial } r = \frac{(\overline{X}_p - \overline{X}_q)pq}{\sigma \times .3989\, h}$$

where

$\overline{X}_p$ = the mean value of the p category[1]

$\overline{X}_q$ = the mean value of the q category

p = percent of cases in p category

q = percent of cases in q category

σ = standard deviation of combined categories (p and q)

h = height of ordinate of normal curve at a distance from the mean including $\frac{p - q}{2}$ of the area of the curve.

The value h is computed by determining the number of standard deviations from the mean including $\frac{p - q}{2}$ of the area of the normal curve (from the table of the area of the normal curve, page 110) and using that number of standard deviations in the table of the

[1] The various categories of the attribute are assigned quantitative values 1, 2 3, etc . if not in quantitative form.

ordinates of the normal curve to determine the height of the curve (in percent of the maximum ordinate) at that point.

ADDITIONAL BIBLIOGRAPHY*

RIETZ, H. L., (Editor), *Handbook of Mathematical Statistics*, Houghton Mifflin Co , New York, 1924

* For readings in standard Statistics textbooks see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents

CHAPTER XII

THE NORMAL CURVE

When events must occur in one of two ways, and there are many such events, they *tend* to be equally divided into two groups; the first consisting of favorable (or desired) occurrences, the second of unfavorable (or non-desired) occurrences. The desired result may be determined empirically by counting and recording the results.

The chances are even that whenever a coin is flipped a "head" will appear. Careful analysis of this somewhat obvious statement will lead to a better understanding of the nature of probability and in turn of the theory of sampling. Since only a "head" or a "tail" can appear, the probability of success in the appearance of the desired face can be stated as one half.

$$p = \frac{1}{2}$$

$$p = \frac{a}{N}$$

where:

a = number of ways in which a favorable outcome can appear.

N = total number of possible events.

p = probability of success.

In general if an event can happen in a ways and not happen in b ways, the probability of its occurrence is

$$p = \frac{a}{a+b} = \frac{a}{N}$$

where

$$a + b = N$$

The probability of failure is therefore:

$$q = \frac{b}{N}$$

where:

b = the possible number of unfavorable results.

q = probability of failure.

If a coin is tossed one hundred times for "heads" or "tails" the probability of either result for each toss is $\frac{1}{2}$. This ratio of probability may be expressed for the group as $(\frac{1}{2})(100)$.

The probability of occurrence of one or another of a series of *mutually exclusive* events may be secured by adding the probabilities of the individual occurrence. To determine the probability

of occurrence of *all* of a series of events the probabilities are multiplied. Thus if a card is drawn from a deck of cards, the probability of either drawing a spade *or* a club is

$$\frac{1}{4} + \frac{1}{4}$$

The probability of the occurrence of *both* of these events is

$$\frac{1}{4} \times \frac{1}{4}$$

The probability that an event will happen or fail to happen is certainty to which is assigned the value 1.

The probability of the appearance of either a head or a tail in the toss of a coin is certainty

$$(\tfrac{1}{2}\,h + \tfrac{1}{2}\,t) = 1$$

In similar manner for a toss of two coins the probability of appearance of heads and tails may be determined from

$$(\tfrac{1}{2}\,h + \tfrac{1}{2}\,t)^2$$

or

$$\tfrac{1}{4}\,h^2 + \tfrac{1}{2}\,ht + \tfrac{1}{4}\,t^2$$

or

probability of all "heads" $= \frac{1}{4}$

probability of all "tails" $= \frac{1}{4}$

probability of a "head" and a "tail" $= \frac{1}{2}$

More generally the probability of occurrence of an event a given number of times in N trials is expressed by[1]

$$N\,(p + q)^n$$

Figure 24 graphically presents the distribution of the number of heads appearing theoretically in tosses of ten coins. As n is increased the number of points to be plotted will increase and the curve will become smoother and smoother ultimately approaching in form the "normal" or Gaussian[2] distribution as seen in Figure 25.

This type of curve frequently results when data exhibiting chance variation are plotted.

The mean of such a distribution may be obtained as follows:

where:

$$\overline{X} = Np$$

N = number of trials

p = probability of success

The standard deviation:

$$\sigma = \sqrt{Npq}$$

where:

q = the probability of failure

[1] The resulting distribution is referred to as the Binomial distribution. It is to be noted that this is a discrete distribution.

[2] The formula for the Gaussian distribution is:

$$y = \frac{N}{\sigma\sqrt{2\pi}}\, e^{\frac{-x^2}{2\sigma^2}}$$

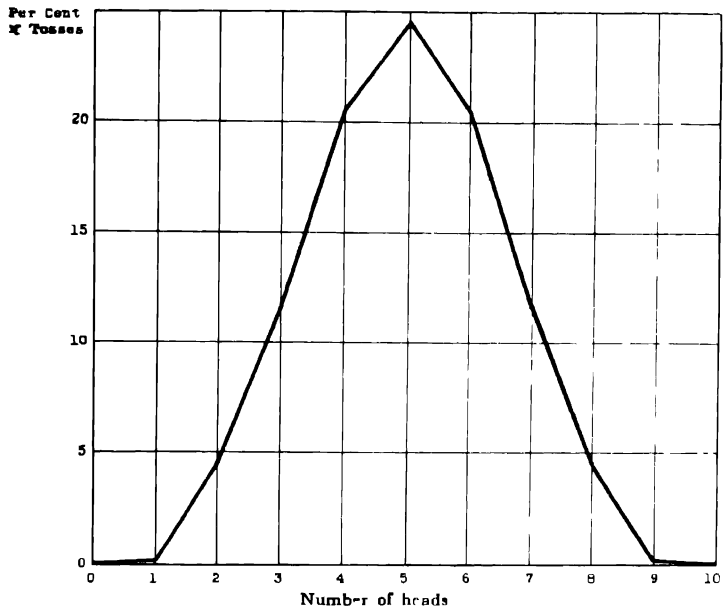


Fig. 24—Theoretical Distribution of Number of "Heads" Appearing in Tosses of 10 Coins.

The distribution is the *normal* or **Gaussian distribution** previously referred to. The standard deviation shown here will bear the same relation to the distribution as outlined previously (see page 39).

Number of Standard Deviations (Measured plus and minus from the mean)	Percent of Cases Included
.6745 σ	50%
1.0000 σ	68.26%
2.0000 σ	95.46%
3.000 σ	99.73%

Generalization of Curves

When only a limited number of items are available the frequency distribution compiled from them is generally irregular in form. If the number of items is increased sharply the distribution will show a tendency to eliminate irregularities and be smoothed.

The sample is frequently used to generalize about the underlying data (technically the **population** or **universe**) and therefore it is often desirable to smooth the irregular curve obtained from the sample. By **smoothing**, the curve of the sample is put into the more general form of the underlying data, or in other words into

an "ideal" distribution. The ideal distribution represents the distribution which would appear if an infinite number of cases were used rather than a sample.

Where it is believed that the underlying data can be best described by means of a normal curve,[1] data can be smoothed by this curve by use of special tables of the ordinates and area of the normal curve (see pages 110-111)[2].

Area Method of Fitting Normal Curve

In a normal distribution the percentage of the cases (area under the curve) included within any number of standard deviations measured from the mean can be determined from specially prepared tables of areas of the normal curve.

If a distance equal to one standard deviation is measured in both directions from the mean it will include 68.26% of the cases, two standard deviations will equal 95.46% of the cases, etc. (see p. 105). In a similar fashion the percentage of cases included within a distance equal to any number of standard deviations but measured in only one direction from the mean will give one half of the percentages shown above.

The mean of the normal curve is located at the center of the distribution. If a distance from the mean to any given point on the X axis is determined *in terms of standard deviations*, the area included within this distance may now be determined by reference to table 31 (page 106). For example, in figure 25 the distance from the mean \$50 to point b \$62 is \$12. The standard deviation of the distribution is given as \$8. Thus the distance from the mean to this point in terms of standard deviations is 1.5 standard deviations (\$12 ÷ \$8). By reference to figure 25 it will be seen that 43.32% of the cases are included between the mean and \$62. If another point on the X axis is now selected, say \$64, it can be found that since this point is 1.75 standard deviations away from the mean $\left(\frac{\$64 - \$50}{8}\right)$, 45.99% of the cases will be included within these limits. Since there are 45.99% of the cases between \$64 and the mean and 43.32% of the cases between \$62 and the mean, 2.67% (45.99% − 43.32%) of the frequencies must occur between \$64 and \$62. As there are 4000 cases in all, 106.8 of them are between \$62 and \$64 in value.

In a similar manner the percentage of cases included within the limits of any one class interval may now be determined, and the number of cases or the theoretical frequency can be found by

[1] See Chapter XV for methods of determining the type of underlying curve.

[2] Various other types of curves may be used to smooth a frequency distribution. Two groups of curves frequently used for this purpose are the Pearson system of curves and the Gram Charlier series.

Since the technique of fitting the curves is beyond the scope of this book, the reader is referred to Elderton, W. P., *Frequency Curves and Correlation* for the Pearson curve, and Camp, B. W., *Mathematical Part of Elementary Statistics* for the Gram Charlier Group.

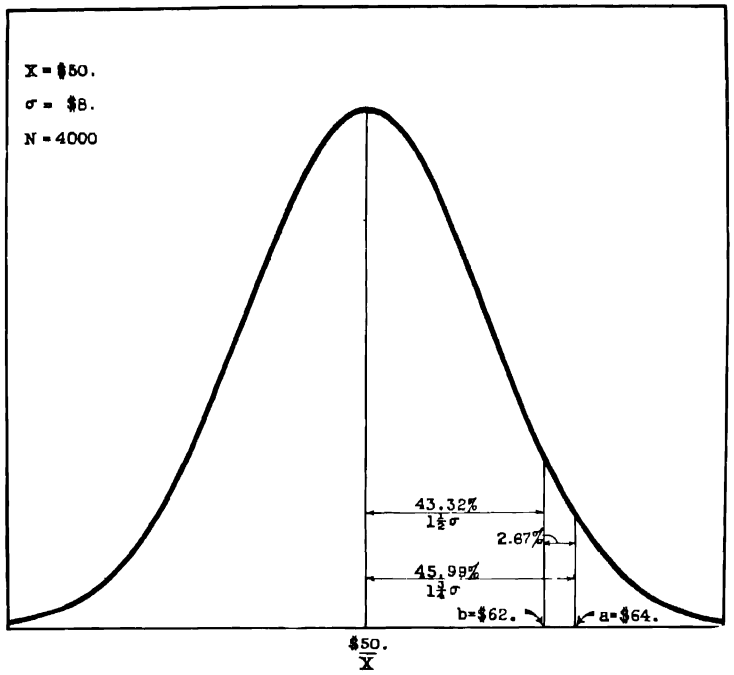


Fig. 25

applying that percentage to the total number of cases. If the theoretical frequencies are then plotted, the result will be a normal curve.

The determination of the theoretical frequency (for one of the class intervals) in the distribution of the variations in the thickness of 600 brass washers manufactured by the A. B. C. Co. as shown below may be used as an example of this procedure. The fitting of this curve will make it possible to determine the variation to be expected in large numbers of these washers as manufactured by this company.

The following measures were computed from the sample:

$$\bar{X} = .0202 \text{ inches}$$
$$\sigma = .00085 \text{ inches}$$
$$N = 600$$

The third class interval has a lower limit of .0188 and an upper limit of .01919. The distance between the lower limit and the mean (.0202 − .0188) is .0014. Since the standard deviation is .00085 inches, in terms of standard deviations, the distance is 1.66 standard deviations. Reference to the area table indicates that 45.15% of the cases are included within this distance. Be-

Table 29—Fitting of Normal Curve Area Method

Variation of Thickness in 600 Brass Washers Manufactured by the ABC Co.

(Hypothetical Data*)

(1) Thickness (In Inches)	(2) Mid-Points	(3) Number of Washers (f)	(4) Deviation of Class Limit From Mean (x)	(5) Column (4) In Terms Of Standard Deviations $\left(\frac{x}{\sigma}\right)$	(6) Percent of Area Between Class Limit and Mean	(7) Percent of Area In Class Interval	(8) Theoretical Frequency f
0180-.01839	.0182	6	− .0022	− 2.61	49 55%	1 21%	7.3
.0184-.01879	.0186	30	− .0018	− 2.13	48.34	3.19	19 1
0188-.01919	.0190	42	− .0014	− 1 66	45.15	7.05	42 3
.0192-.01959	.0194	66	− .0010	− 1 18	38.10	11 98	71.9
.0196-.01999	.0196	94	− .0006	.71	26.12	16 64	99 8
.0200- 02039	.0202	120	− .0002 .0002	.24 .24	9 48 9.48	18.96	113.8
.0204-.02079	.0206	102	.0006	.71	26.12	16 64	99 8
.0208-.02119	.0210	60	.0010	− 1.18	38 10	11.98	71.9
.0212-.02159	.0214	54	0014	− 1 66	45.15	7 05	42 3
.0216-.02199	.0218	14	0018	2.13	48 34	3 19	19.1
.0220-.02239	.0222	12	0022	2.61	49.55	1.21	7.3
		600					

* Hypothetical data based on smaller distribution given by W. A. Shewhart, *Economic Control of Quality of Manufactured Product.*

tween the lower limit of the next class interval and the mean (.0202 − .0192 = .0010) there is a distance equal to 1.18 standard deviations. The area table indicates that 38.10% of the cases are included within this distance. It can then be seen that there must be 7.05% of the cases between .0188 and .01919 or between the upper and lower limits of group three. Since the total frequency is given as 600 the theoretical frequency for this particular class interval will be (7.05% of 600). The frequency is then plotted at the mid-point of the group. The same process is repeated for all other class intervals as shown in table 29.

Fitting the Normal Curve—Ordinate Method

The normal curve may also be fitted by reference to a table of ordinates of the probability curve (p. 115). The table gives the ordinates of the normal curve at any distance (in terms of standard deviations) from the mean as a percent of the maximum ordinate. The maximum ordinate occurs at the center of the distribution.

The formula for the normal curve is

$$Y = Y_0 e^{\frac{-x^2}{2\sigma^2}}$$

where[1]

Y_o, the maximum ordinate $\dfrac{N}{\sigma\sqrt{2\pi}}$ $\dfrac{N}{2.506628\ \sigma}$

The value of the maximum ordinate (Y_o) for the distribution is computed below:[1]

σ (class interval units) = 2 109

$$Y_o \quad \frac{N}{2\ 506628\ \sigma} \quad \frac{600}{2.109(2\ 506628)} \quad 113\ 5$$

For the midpoint of the first group which is 2.37 standard deviations (.0020 inches) from the mean we find that an ordinate erected at this point would be 6.03% of the maximum ordinate or 6 84. The same procedure may be used on the other class intervals as shown in table 30.

Table 30—Fitting of Normal Curve—Ordinate Method
Variation of Thickness in 600 Brass Washers Manufactured by the ABC Co.

(1) Thickness (In Inches)	(2) Mid-Points	(3) Number of Washers (f_o)	(4) Deviation of Mid-Point from Mean (x)	Column 4 In Terms of Standard Deviations $\left(\frac{x}{\sigma}\right)$	Percent of Maximum Ordinate (From Ordinate Table)	Theoretical Frequency (f)
0180– 01839	0182	6	0020	2 37	6 03%	6 84
0184– 01879	0186	30	0016	1 90	16 45	18 67
0188– 01919	0190	42	0012	1 42	36 49	41 42
0192– 01959	0194	66	0008	95	63 68	72 28
0196– 01999	0196	94	0004	47	89 54	101 62
0200– 02039	0202	120	0000	00	100 00	113 50
0204– 02079	0206	102	0004	47	89 54	101 628
0208– 02119	0210	60	0008	95	63 68	72 28
0212– 02159	0214	54	0012	1 42	36 49	41 42
0216– 02199	0218	14	0016	1 90	16 45	18 67
0220– 02239	0222	12	0020	2 37	6 03	6 84
		600				

Testing the Goodness of Fit—The Chi Square Test

A test to determine the goodness of fit of the actual data to the theoretical distribution has been devised by Karl Pearson

The test involves the calculation of χ^2 (chi square)*

where $$\chi^2 = \Sigma\left(\frac{(f_o - f)^2}{f}\right)$$

f_o = the observed or actual frequencies

f = the theoretical frequencies

For the problem outlined above chi square may be calculated as in table 29a [2]

[1] In the application of this formula the standard deviation in class interval units, not original units, must be used

[2] If the value of f for any class interval is very small, several groups must be combined in order not to obtain disproportionate values of χ^2 See Table 29a

* The symbol χ is the Greek small letter chi

Table 31—Normal Curve Area Table

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0159	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0 6	.2257	.2291	.2324	2357	.2389	.2422	.2454	.2486	.2518	.2549
0 7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3380
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3718	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	3997	.4015
1.3	.4032	.4049	.4066	.4083	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4430	4441
1 6	.4452	.4463	.4474	.4485	.4495	.4505	.4515	.4525	.4535	.4545
1.7	4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	4750	.4758	.4762	.4767
2.0	.4773	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4865	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4980	.4980	4981
2.9	.4981	.4982	.4983	.4984	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.4987	.4987	.4988	.4988	.4988	.4989	.4989	.4989	.4990
3.1	.49903	4991	.4991	.4991	.4992	.4992	.4992	4992	.4993	.4993

Table 29a—The Chi Square Test for Goodness of Fit—Data of Table 29

(1) Thickness (In Inches)	(2) Number of Washers (f_o)	(3) Theoretical Frequency (f)	(4) $(f_o - f)$	(5) $(f_o - f)^2$	(6) $\frac{(f_o - f)^2}{f}$
.0180–.01839	6	7.3	9.6	92.16	3.491
.0184–.01879	30	19.1			
.0188–.01919	42	42.3	.3	.09	.007
.0192–.01959	66	71.9	− 5.9	34.81	.484
.0196–.01999	94	99.8	− 5.8	33.64	.337
.0200–.02039	120	113.8	6.2	38.44	.338
.0204–.02079	102	99.8	2.4	5.76	.058
.0208–.02119	60	71.9	− 11.9	141.61	1.970
.0212–.02159	54	42.3	12.3	151.29	3.628
.0216–.02199	14	19.1	− .4	.16	.001
.0220–.02239	12	7.3			
					$\chi^2 = 10.314$

Table 32—Normal Curve Ordinates
(As decimal of maximum ordinate)

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	1.00000	.99995	.99980	.99955	.99920	.99875	.99820	.99755	.99685	.99596
0.1	.99501	.99396	.99283	.99158	.99025	.98881	.98728	.98565	.98393	.98211
0.2	.98020	.97819	.97609	.97390	.97161	.96923	.96676	.96420	.96156	.95882
0.3	.95600	.95309	.95010	.94702	94387	.94055	.93723	.93382	.93024	.92677
0.4	.92312	.91399	.91558	.91169	90774	.90371	.89961	.89543	.89119	.88688
0.5	.88250	.87805	.87353	.86896	.86432	.85962	.85488	.85006	.84519	.84060
0 6	.83527	.83023	.83514	.82010	.81481	.80957	.80429	.79896	.79459	.78817
0.7	.78270	.77721	.77167	.76610	.76048	75484	.74916	.74342	.73769	.73193
0.8	.72615	.72033	.71448	.70861	.70272	.69681	.69087	.68493	.67896	.67298
0.9	.66689	.66097	.65494	64891	.64287	.63683	.63077	.62472	.61865	.61259
1 0	.60653	.60047	.59440	.58834	.58228	.57623	.57017	56414	.55810	.55209
1 1	.54607	.54007	.53409	.52812	.52214	.51620	.51027	.50437	.49848	.49260
1.2	.48675	.48092	.47511	.46933	.46357	.45793	.45212	.44644	.44078	.43516
1.3	.42956	.42399	.41845	.41294	.40747	.40202	.39661	.39123	.38569	.38058
1.4	.37531	.37007	36487	35971	35459	34950	.34445	33944	.33447	.32954
1.5	.32465	.31980	.31500	.31023	.30550	30082	.29618	.29158	.28702	.28251
1.6	.27804	.27361	.26923	.26489	.26059	25634	.25213	.24797	.24385	.23978
1.7	.23575	.23176	.22782	.22392	.22008	.21627	.21251	.20879	.20511	.20148
1.8	.19790	.19436	.19086	.18741	.18400	18064	.17732	.17904	.17081	.16762
1.9	.16448	.16137	.15831	.15530	.15232	.14939	.14650	.14364	14083	.13806
2.0	.13534	.13265	.13000	.12740	.12483	.12230	.11981	.11737	.11496	.11259
2.1	.11025	.10795	.10570	.10347	.10129	.09914	.09702	.09495	.09290	.09090
2.2	.08892	.08698	.08507	.08320	.08136	.07956	.07778	.07604	.07433	.07265
2.3	.07100	.06939	.06780	.06624	.06471	.06321	.06174	.06029	.05888	.05750
2.4	.05614	.05481	.05350	.05222	.05096	.04973	.04852	.04737	.04618	.04505
2.5	.04394	.04285	.04179	04074	.03972	.03873	.03775	.03680	.03586	.03494
2 6	.03405	.03317	.03232	.03148	.03066	.02986	.02908	.02831	.02757	.02684
2 7	.02612	.02542	.02474	02408	.02343	.02280	.02218	.02157	.02098	.02040
2.8	.01984	.01929	.01876	.01823	.01772	.01723	.01674	.01627	.01581	.01536
2.9	.01492	.01449	.01408	.01367	.01328	01288	.01252	.01215	.01179	.01145
3.0	.01111	.00819	.00598	.00432	00309	.00219	.00153	.00106	.00073	.00050
4 0	.00034	.00022	00015	00010	00006	00004	.00003	.00002	.00001	.00001

By reference to a set of χ^2 tables[1] chi square may be evaluated.

In tables of χ^2, N equals the number of class intervals. The value indicated in the table (p) is the probability of obtaining a fit, due to chance, *as poor as or worse* than the one obtained. If this probability is small the likelihood that the disparities between the observed and actual data are due to chance is small.[2]

The value of χ^2 in the problem above for $N = 9$ class intervals indicates a value for p of more than 30. In other words there are more than 30 chances out of 100 that the fit obtained would be as bad or as worse than the one shown. 30 chances out of 100 the variations or departures from normality occurring in the sample might be as bad as or worse than those actually occurring, due to chance fluctuations with no real departure from normality.

[1] An extended set of these tables is to be found in Pearson's *Tables for Statisticians and Biometricians.*

[2] Generally, if the indicated value of p is less than some specified value, usually .05 or .01, the discrepancies are accepted as too large to be accidental. The accepted limit value of p (usually .05 or .01) is referred to as the fiducial point.

The chi square test may be used to test a large variety of hypotheses in many fields of comparing the expected results (frequencies) based upon the hypothesis to be tested and the actual results obtained by securing observations. If the chi square test demonstrates that the disparity between the actual and the expected frequencies is too large to be ascribable to chance (if p is less than the selected fiducial limit of .01 or .05), the hypothesis may be said to be false.

Section of χ Table

N	n $(N-1)$	$p = .99$	$p = .05$	$p = .01$
2	1	.00016	3.84	6.64
3	2	.020	5.99	9.21
4	3	.115	7.82	11.34
5	4	.297	9.49	13.28
6	5	.554	11.07	15.09
7	6	.872	12.59	16.81
8	7	1.239	14.07	18.48
9	8	1.646	15.51	20.09
10	9	2.088	16.92	21.67
11	10	2.558	18.31	23.21
12	11	3.053	19.68	24.73
13	12	3.571	21.03	26.22
14	13	4.107	22.36	27.69
15	14	4.660	23.69	29.14
16	15	5.229	25.00	30.58

ADDITIONAL BIBLIOGRAPHY*

BOWLEY, ARTHUR L., *Elements of Statistics*, pp. 259–311. P. S. King & Son, London, 1901.

FISHER, R. A., *Statistical Methods for Research Workers*, pp. 43–53; 64–70. Oliver & Boyd, Edinburgh, 1932.

ODELL, C. W., *Educational Statistics*, pp. 305–315. Century Co., New York, 1925.

RIETZ, H. L., (Editor), *Handbook of Mathematical Statistics*, pp. 82–91; 92–119. Houghton Mifflin Co., New York, 1924.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER XIII

THEORY OF SAMPLING

The Sample

Statistical technique as applied to a given mass of data enables us to analyze that data. However, where the mass of data is too great to be handled in its entirety, given samples of the data subjected to the same technique enable us to generalize about the mass from which the sample is drawn.

If the results are confined to the cases studied they may be used for descriptive purpose. A number of problems, however, must be solved before the results can be generalized and applied to the larger number of cases not included in the original study.

Much time, energy and money would be needed to make a comprehensive analysis of a great mass of statistical data. Consequently there is every incentive to resort to study of part or parts of the data. This process is known as sampling. The validity of the results obtained depends upon the fairness of the sample and the technique employed in studying that sample.

A few cases in which the sampling method is indicated are outlined below.

In the physical sciences it frequently becomes impossible to obtain further data and sampling must be adopted—i.e., as when an experiment cannot be performed beyond a given number of times. Also, the results of a scientific experiment repeated ten times can be used as a generalization of the results which might logically be assumed to be obtainable if the experiment were performed an infinite number of times.

To obtain an average Intelligence Quotient for third grade public school students the use of any other method than sampling would mean the very expensive accumulation of an enormous mass of data.

Again, let us say it is desired to obtain the average price of bread in New York City. Obviously the cost factor and the time required for a complete survey of the city's thousands of bakeshops and grocery stores would be prohibitive.

In the last instance cited if prices were obtained from chain stores only the sample would be prejudiced. To secure *representative* data it would be necessary to obtain sample prices from all the varied types of stores; from, in technical term, the entire **population**[1]. The requirements for a representative sample can be summed up into:

[1] The population may be defined as the entire data from which the sample was drawn if all of it were available.

1. The sample must be selected without bias or prejudice.
2. The components of the sample must be completely independent of one another.
3. There should be no underlying differences between areas from which the data are selected.
4. Conditions must be the same for all items constituting the sample.

The limited sample is most generally used to describe the larger population or group of data from which the sample was taken.

When the measures computed from a sample are used to characterize the population, it is necessary to estimate the reliability of the measures, in other words the degree of error to which the generalization may be subject.

Samples may be drawn from the underlying population in several different ways. The conditions outlined above are descriptive of the *random sample*. The values composing a sample of this type are drawn entirely at random from the population.

However, it is often desirable to segregate a heterogeneous population into homogeneous sub groups and to draw from each sub group at random. This process results in the *stratified* sample. A survey of the number of rooms in homes in a given city by resort to the sampling method may be more effectively secured by drawing a random sample from uniform areas of the city. Thus a random sample of a low income area in the same proportion that the population of that area bears to the total population might be combined with random samples drawn in the proper proportion from other income areas.

The *purposive sample* represents a deliberate selection of a sample manipulated by the statistician in such a fashion as to obtain a representative cross section of the population.

Measures of Reliability and Significance—Standard Errors

In the bread price problem outlined above, let us assume that one thousand investigators are sent out each to obtain a sample of prices for a given size and type of bread from 1000 stores. If the average prices for the samples were then arranged in the form of a frequency distribution they would be found to tend toward a "normal" distribution.

The average of the means of the 10,000 samples of 1000 prices each would undoubtedly result in a figure either the same as or very near the true mean of the underlying data. When the curve is idealized as though an infinite number of cases had been considered, the "true" mean for bread prices in New York could be obtained by computing the average of the distribution.

From the hypothetical normal distribution shown above, it can be seen that the means of some of the samples were quite a

distance from the "true" mean of the population. If the distance of the mean furthest away can be obtained the greatest possible error will be known. It has already been shown that 99.7% of the cases will be included within a distance of 3 standard deviations from the mean (see p. 38). If the value of the standard deviation of this distribution (of the means of the samples) can be obtained, in 99.7 chances out of 100 no error (difference between sample mean and "true" mean) will occur larger than 3 times the value of the standard deviation.

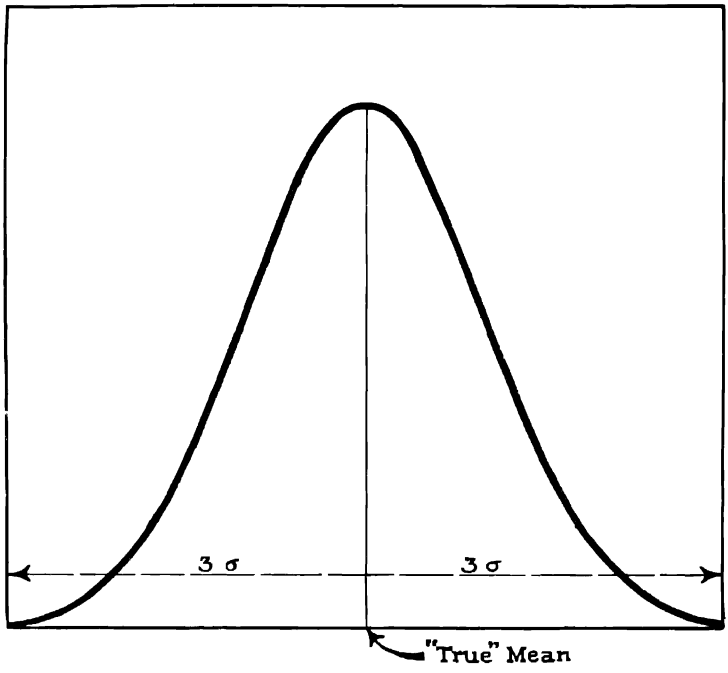


Fig. 26—Theoretical Distribution of Means of a Large Number of Samples Drawn at Random from a Given "Population."

The standard deviation of a distribution of means or any other statistical measure computed from samples is termed the **standard error of the mean** ($\sigma_{\bar{x}}$) or the standard error of any other statistical measure.

The error which will not be exceeded by 50% of the cases is known as the **probable error**[1]. It is equal to .6745 times the standard error.

[1] Although widely used the probable error is of comparatively little value. It can be interpreted as meaning that if another sample were drawn of the same number of items the chances are even that a discrepancy between the sample and true mean larger than the probable error would not exist.

Extensive use of the **probable error**, notably by American statisticians, gives it a fictitious value far beyond its real worth as compared with the **standard error**.[1]

It is obvious that the greater the number of cases included in the sample the smaller the error to be expected. Therefore the standard deviation (error) of the theoretical distribution of means (or other measure) computed from samples will be smaller. In turn the standard error will vary inversely with the number of cases included in the sample.

When the limits of variation in the **population** or original data are great, for example a range of \$1 to \$1,000,000., a greater "error" is to be expected in a measure computed from the sample when the range of values in the population is small; i.e., from \$1 to \$10. The standard deviation is the measure of the spread of data of the population. The standard deviation of the sample is commonly used as an estimate of the standard deviation of the population. It follows that the standard error of a measure computed from a sample will vary directly with the standard deviation of the population from which the sample was drawn.

The formula for the standard error of the mean (the standard deviation of the distribution of the means of samples) is

$$\sigma_{\bar{x}} \qquad \sqrt{N}$$

where

σ = standard deviation of sample.

The probable error of the mean is therefore

$$\text{P.E.}_{\bar{x}} = .6745\,\frac{\sigma}{\sqrt{N}}$$

Reliability can be evaluated if the average price of bread is computed from one sample of prices obtained from 1000 stores. If the following results were obtained from the sample:

$$\overline{X} = \$.10$$
$$\sigma = \$.01$$

the standard error of the mean would be

$$\sigma_{\bar{x}} = \frac{\$.01}{\sqrt{1000}} = \$.00032$$

As shown above there are 99.7 chances out of 100 that the mean computed from a random sample of 1000 cases will not be further away than three standard errors of the mean from the true average

[1] R. A. Fisher points out that "the common use of the probable error is its only recommendation . . . when any critical test is required, the deviation must be expressed in terms of the standard error."* It has also been pointed out that the standard error has been commonly used by European statisticians in preference to the probable error.

* Fisher, R. A., *Statistical Methods for Research Workers*, 1928, p. 46.

price. It can now be assumed that the computed average of $.10 will not be more than $.00096 away from the "true" average (99.7 chances out of 100).

The probability of occurrence and the odds against the occurrence of an error as great as the given number of standard errors as given by Pearl[1] are shown in table 32a.

The formula for the standard error of the mean, shown above, is for the mean of a random sample. The formula for the standard error of the mean of a stratified sample is[2]

$$\sigma_{\bar{x}_s}^2 = \frac{\sigma^2}{N} - \frac{\sigma_m^2}{N}$$

where

σ_m = standard deviation of the means of each of the strata comprising the sample.

In computing the standard deviation of the means of the strata, the deviations of the means of each strata from the mean of all the values must be weighted in proportion to the number of items in each strata.

It will be seen from this formula that the standard error of the mean of a stratified sample will always be equal to or less than that of the mean of a random sample and the mean of a stratified sample therefor equally or more reliable than the mean of a random sample.

The formulas discussed above apply to the means of samples drawn from infinitely large populations or means of samples which are very small in comparison to the size of the underlying population. When the population is finite in size, particularly when the size of the sample is appreciable in proportion to the size of the universe, the standard error of the mean may be computed from the formula[3]

$$\sigma_{\bar{x}_p} = \sigma_{\bar{x}} \sqrt{1 - \frac{n}{N}}$$

or

$$\sigma_{\bar{x}_p} = \frac{\sigma}{\sqrt{N}} \sqrt{1 - \frac{n}{N}}$$

where

n = number of items in sample

N = number of items in population the standard error of the mean of a random sample as shown above.

[1] Pearl, Raymond, *Medical Biometry and Statistics*, W. B. Saunders, Philadelphia, 1930

[2] Yule, G. U., and Kendall, M. G., *An Introduction to the Theory of Statistics*, London, 1937, p. 389.

[3] Bowley, Arthur L., *Elements of Statistics*, Sixth Edition, 1937, pp. 332-333. This formula is given by Richardson as.

$$\sigma_{\bar{x}_p} = \sqrt{\frac{N-n}{N-1}}\,\sigma_{\bar{x}}$$

Richardson, C. H., *An Introduction to Statistical Analysis*, Harcourt, Brace & Co., New York 1935, p. 259.

In a similar fashion the standard errors of other statistical measures may be computed. The formulae for some of these standard errors are shown below (page 119).

This list represents merely a small number of the standard error formulae. For a more complete listing see the list of formulas at the rear of this book and *Dunlap, J. W. and Kurtz, A. K., Handbook of Statistical Nomographs, Tables and Formulae.*

Table 32a—The Probability of Occurrence of Statistical Deviations of Different Magnitudes Relative to the Standard Error

Number of Standard Errors	Probability of occurence a deviation as great as or greater than designated number of Standard Errors	Odds against the occurrence of a deviation as great as or greater than the designated number of Standard Errors
0.67449	50.00%	1.00 to 1
0.7	48.39	1.07 to 1
0.8	42.37	1.36 to 1
0.9	36.81	1.72 to 1
1.0	31.73	2.15 to 1
1.1	27.13	2.69 to 1
1.2	23.01	3.35 to 1
1.3	19.36	4.17 to 1
1.4	16.15	5.19 to 1
1.5	13.36	6.48 to 1
1.6	10.96	8.12 to 1
1.7	8.91	10.22 to 1
1.8	7.19	12.92 to 1
1.9	5.74	16.41 to 1
2.0	4.55	20.98 to 1
2.1	3.57	26.99 to 1
2.2	2.78	34.96 to 1
2.3	2.14	45.62 to 1
2.4	1.64	60.00 to 1
2.5	1.24	79.52 to 1
2.6	.932	106.3 to 1
2.7	.693	143.2 to 1
2.8	.511	194.7 to 1
2.9	.373	267.0 to 1
3.0	.270	369.4 to 1
3.1	.194	515.7 to 1
3.2	.137	726.7 to 1
3.3	.0967	1,033 to 1
3.4	.0674	1,483 to 1
3.5	.0465	2,149 to 1
3.6	.0318	3,142 to 1
3.7	.0216	4,637 to 1
3.8	.0145	6,915 to 1
3.9	.00962	10,390 to 1
4.0	.00634	15,770 to 1
5.0	.0000573	1,744,000 to 1
6.0	.00000020	500,000,000 to 1
7.0	.00000000026	400,000,000,000 to 1

Measure	Standard Error Formula	Probable Error Formula
Mean ($\bar{X}$)	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$	$\text{P.E.}_{\bar{x}} = .6745\frac{\sigma}{\sqrt{N}}$
Median	$\sigma_{mdn} = 1.2533\frac{\sigma}{\sqrt{N}}$	$\text{P.E.}_{mdn} = .84535\frac{\sigma}{\sqrt{N}}$
Standard Deviation (σ) (for sample from normally distributed universe See p 198)	$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}}$	$\text{P.E.}_\sigma = .6745\frac{\sigma}{\sqrt{2N}}$
Mean Deviation	$\sigma_{M.D.} = .6028\frac{\sigma}{\sqrt{N}}$	$\text{P.E.}_{M.D.} = .4066\frac{\sigma}{\sqrt{N}}$
Coefficient of Variation (V)	$\sigma_v = \frac{V}{\sqrt{2N}}\sqrt{1+\frac{2(V)^2}{(10)^4}}$	$\text{P.E.}_v = .6745\frac{V}{\sqrt{2N}}\sqrt{1+\frac{2(V)^2}{(10)^4}}$
Coefficient of Correlation (r)	$\sigma_r = \frac{1-r^2}{\sqrt{N}}$	$\text{P.E.}_r = .6745\frac{1-r^2}{\sqrt{N}}$
Coefficient of Rank Correlation (ρ)	$\sigma = \frac{1}{\sqrt{N-1}}$	$\text{P.E.}_\rho = .6745\frac{1}{\sqrt{N-1}}$
Multiple Correlation Coefficient ($R_{1\,23\ldots n}$)	$\sigma_{R1.23\ldots n} = \frac{1-R^2_{1.23\ldots n}}{\sqrt{N}}$	$\text{P.E.}_{R1.23\ldots n} = .6745\frac{1-R^2_{1.23\ldots n}}{\sqrt{N}}$
Partial Correlation Coefficient	$\sigma_{r12.34\ldots n} = \frac{1-r^2_{12.34\ldots n}}{\sqrt{N}}$	$\text{P.E.}_{r12.34\ldots n} = .6745\frac{1-r^2_{12.34\ldots n}}{\sqrt{N}}$

Significance of the Difference Between Two Means

Very frequently it is desirable to test the means of two samples to determine whether there is any significant difference between them, or whether the difference, if any, is merely due to chance.

In scientific fields it is customary to use a "control" when carrying out an experiment. This control, to which the new technique is not applied, is used as a basis for comparison. It is then essential to determine whether the measured results for the "control" group differs significantly from the experimental group.

As an instance, if a new technique for teaching spelling is subjected to experimental test two groups of pupils are used. The new technique as applied to group 1 and group 2 is used solely as a control. The results (grades on examinations, etc.) are then tested to determine the significance of the difference between the mean grades for the two groups.

If two samples of any data are drawn from a given population undoubtedly there will be a difference between the means of the samples, a difference due solely to chance variations in the selection of items.

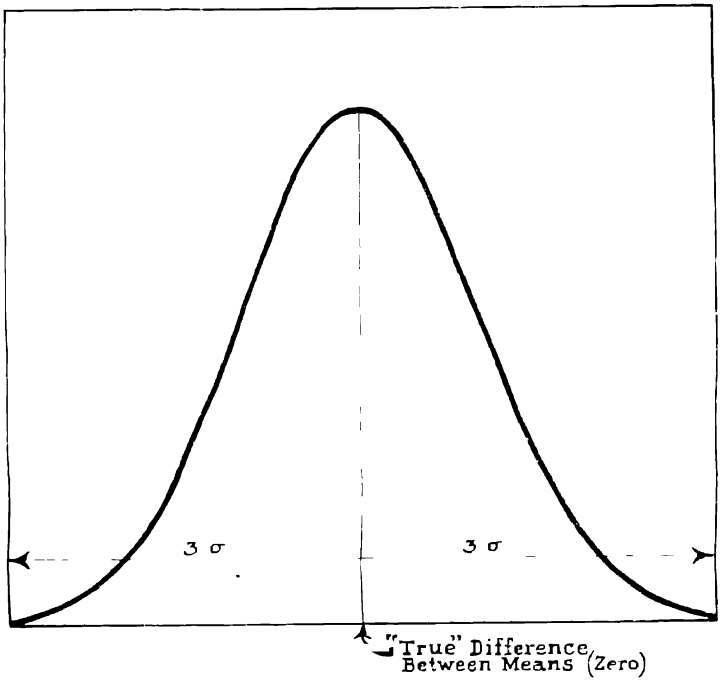


Fig. 27—Theoretical Distribution of Difference Between Means of a Large Number of Pairs of Samples Drawn at Random from a Given "Population."

If a very large number of these pairs of samples are drawn from the population, and if the difference between the means of the pairs is arranged in the form of a frequency distribution, the resulting distribution will be normal. The true difference is in reality zero, since all these pairs of samples are drawn from the same population and only *chance or accidental* differences arise between the samples. If all the differences determined from a very large or more exactly an infinite number of pairs are averaged (signs retained), the true difference (zero) will result.

The situation may then be shown by a normal curve representing a distribution of the differences between the means of an infinite number of pairs of samples. The mean of this distribution would then be zero, the true difference, and positive and negative differences due to chance would arise about this mean.

It is obvious from the information previously given about the normal curve that, 99.7 chances out of 100, no difference larger than 3 standard deviations of this distribution of differences (three standard errors of the difference between the two means) would arise. If, therefore, the actual difference is larger than 3 standard errors of the difference between the means (or, in other words, if the probability of such a difference due to chance is very small) it can be said that the difference is significant and not due to chance.

The standard error of the difference between two means (standard deviation of the theoretical distribution of differences between means of samples) can be obtained from the following:[1]

$$\sigma_D = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}$$

$$= \sqrt{\frac{\sigma^2_1}{N_1} + \frac{\sigma^2_2}{N_2}}$$

where

σ_1 = standard deviation of first sample
σ_2 = standard deviation of the second sample
N_1 = number of items in first sample
N_2 = number of items in second sample.

As a result of a time study a new method was outlined for a certain operation in a factory. The average time on fifty trials for the operation using the old method was 17.5 seconds with a standard deviation of 1.5 seconds. After learning the new method the workmen were again timed with a resulting average time for fifty trials of 15 seconds and a standard deviation of 1.2 seconds. It is now possible by means of the technique outlined above to

[1] This is based on the assumption that the two samples from which the means were computed are uncorrelated.

determine whether the difference in average time of 2.5 seconds is significant or merely due to chance.

$$\sigma_D = \sqrt{\frac{(1.5)^2}{50} + \frac{(1.2)^2}{50}} = .27 \text{ seconds}$$

A difference as large as .81 seconds (3 standard errors of the difference between the two means) might (99.7 chances in 100) arise due to chance. Since the actual difference (2.5 seconds) is much larger than this amount (.81 seconds) it is extremely unlikely that it arose due to chance.

The significance of the difference between any two statistical measures computed from two samples may be obtained from:

A. If the two samples are correlated

$$\sigma_D = \sqrt{\sigma^2_{\theta_1} + \sigma^2_{\theta_2} - 2\, r_{12}\, \sigma_{\theta_1}\, \sigma_{\theta_2}}$$

where

σ_{θ_1} is the standard error of any statistic θ computed from sample #1

σ_{θ_2} is the standard error of any statistic computed from sample #2

B. If the samples are uncorrelated

$$\sigma_D = \sqrt{\sigma^2_{\theta_1} + \sigma^2_{\theta_2}}$$

Significance of Difference Between Proportions

If two random samples are drawn and indicate that a given characteristic is in a certain proportion, the difference between the two proportions can be tested to determine whether it is significant or arises out of a sampling fluctuation by use of the formula

$$\sigma_{D\%} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

where

p is the total percentage of occurrence

$q = 1 - p$

N_1 = number in first sample

N_2 = number in second sample

In a study of the effectiveness of slogans it was found that 75.7% of the males questioned recognized a certain slogan while 66.3% of the females questioned recognized the slogan. The above formula may be applied to determine whether there was a significant difference in the percentage of recognition by the two sexes.

Table 32b—Results of Recognition Test of Paris Garter Slogan, "No Metal Can Touch You," as Given to 374 College Students

Sex	Number Recognizing	Percent Recognizing	Number Questioned
Male	209	75.7%	276
Female	65	66.3%	98
Total	274	73.3%	374

From Glick, S., *Commercial Slogans*, Thesis, College of City of New York, 1935.

$$p = 73.3\%$$
$$q = 26.7\%$$
$$\sigma_{D\%} = \sqrt{p\,q\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{(.733)\,(.267)\left(\frac{1}{276} + \frac{1}{98}\right)}$$
$$= .052 = 5.2\%$$

Since the actual difference between the two proportions (75.7% − 66.3% = 9.4%) is 1.81 times the standard error of the difference, there are a little over 7 chances in 100 that the difference is a chance difference due to sampling.

Standard Error of Measurements

A certain degree of variation must be expected when physical measurements are performed. If a distance is measured repeatedly or if a quantity is weighed several times, the results will show a degree of variation.

If the average of the several measurements is taken as the true measurement it must be remembered that this average is a measurement obtained from a sample. It is therefore subject to a sampling error which may be computed. If a measurement is made 10 times it constitutes a sample of 10 measurements drawn from a universe of an infinite number of measurements which may be made.

It has been shown previously that the error of such a mean can be computed through the use of the standard error or probable error of the mean.

For Large Samples ($N > 30$)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

$$\text{P.E}_{\bar{x}} = .6745\,\frac{\sigma}{\sqrt{N}}$$

For Small Samples ($N < 30$) (See page 129)

$$S_{\bar{x}} = \frac{S}{\sqrt{N}}$$

However it is frequently necessary to combine measurements to obtain areas, volumes, etc. In this case the standard error of the resulting value must be obtained.

1. When the individual measurements are combined by addition the standard error of the resulting value, may be obtained from[1]

$$\sigma^2_{\bar{X}_1+\bar{X}_2+\bar{X}_3+\quad .+\bar{X}} = \sigma_{\bar{X}1}^2 + \sigma_{\bar{X}2}^2 + \sigma_{\bar{X}3}^2 + \ldots + \sigma_{\bar{X}}^2$$

To obtain the distance between two points the distance was measured in two separate sections, with the following results, as an average of a number of measurements

Distance #1 = 500 yds.

Distance #2 = 600 yds.

[1] These relationships are true only if the items are mutually independent for when they are not the relationship is

$$\sigma_{\bar{X}1 + \bar{X}_2} = \sqrt{\sigma_{\bar{X}1}^2 + \sigma_{\bar{X}2}^2 + 2r_{12}\sigma_{\bar{X}1}\sigma_{\bar{X}2}}$$

The standard error of the measurements of the first distance $\left(\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\right)$ was 2 yards while that of the second 2 5 yards. The standard error of the entire distance, 500 yards + 600 yards, is

$$\sigma^2_{\bar{X}_1+\bar{X}_2} = 4 + 6.25 = 10.25$$

$$\sigma_{\bar{X}_1+\bar{X}_2} = 3.20 \text{ yards}$$

2. If the measurement is raised to any power (n) the standard error of the resulting value may be obtained from

$$\frac{\sigma_{\bar{X}}^n}{\bar{X}^n} = N\left(\frac{\sigma_{\bar{X}}}{\bar{X}}\right)$$

The measurements of one side of a square result in an average length of 10 feet with a standard error of .05 feet. The standard error of the area can be obtained as follows.

$$\text{Area} = L^2 = 10^2 = 100 \text{ square feet}$$

$$\frac{\sigma_{\bar{X}}^n}{100} = 2\left(\frac{.05}{10}\right) = .1 \text{ square foot}$$

3. The standard error of the product of a series of means the standard errors of which are known, is obtained from:

$$\left(\frac{\sigma_{\bar{X}_1}\ \sigma_{\bar{X}_2}\cdots\sigma^n}{\bar{X}_1\ \bar{X}_2 \ldots \bar{X}_n}\right)^2 = \left(\frac{\sigma_{\bar{X}_1}}{\bar{X}_1}\right)^2 + \left(\frac{\sigma_{\bar{X}_2}}{\bar{X}_2}\right)^2 + \ldots + \left(\frac{\sigma_{\bar{X}_n}}{\bar{X}_n}\right)^2$$

4. The standard error of a quotient can be obtained from

$$\left(\frac{\sigma_{\frac{\bar{X}1}{\bar{X}2}}}{\frac{\bar{X}_1}{\bar{X}_2}}\right)^2 = \left(\frac{\sigma_{\bar{X}_1}}{\bar{X}_1}\right)^2 + \left(\frac{\sigma_{\bar{X}_2}}{\bar{X}_1}\right)^2$$

The standard error of the volume of a cylindrical tank as obtained from a series of measurements of its radius and height is obtained as follows:

$$\text{radius} = 10 \text{ feet (r)}$$
$$\text{height} = 20 \text{ feet (h)}$$
$$V = \pi r^2 h = 3.14159(10)^2(20) = 6283.18$$

and

$$\sigma_r = .1 \text{ feet}$$
$$\sigma_h = .2 \text{ feet}$$

The standard error of r^2 can be obtained from

$$\frac{\sigma_{\bar{X}^n}}{\bar{X}^n} = N\frac{\sigma_{\bar{X}}}{\bar{X}} = \frac{\sigma_{r^2}}{r^2} = 2\frac{\sigma_r}{r} = \left(2\frac{.1}{10}\right)$$

and the standard error of the volume from

$$\left(\frac{\sigma_{\bar{X}_1 \cdot \bar{X}_2}}{\bar{X}_1 \cdot \bar{X}_2}\right)^2 = \left(\frac{\sigma_{\bar{X}_1}}{\bar{X}_1}\right)^2 + \left(\frac{\sigma_{\bar{X}_2}}{\bar{X}_2}\right)^2$$

or in this case where $V = \pi r^2 h$

$$\left(\frac{\sigma_v}{V}\right)^2 = \left(\frac{\sigma_{v^2}}{r^2 h}\right) = \left(2\frac{\sigma_r}{r}\right)^2 + \left(\frac{\sigma_h}{h}\right)^2$$

$$\left(\frac{\sigma_v}{6283.18}\right)^2 = \left(2\frac{1}{10}\right)^2 + \left(\frac{?}{20}\right)^2$$

$$\sigma_v = 140.5 \text{ cubic feet}$$

Significance of Coefficient of Correlation

When a coefficient of correlation is computed it is necessary to determine whether or not the correlation indicated is a real association between the two considered series, or whether the indicated relationship has arisen from the accidental selection of values in the samples.

Although no real association exists between the two series constituting the sample it is possible to obtain a definite value for r when computed from a sample drawn from the universe. It has already been noted that a difference between the means of two samples may make its appearance even when both samples are drawn from the same population. In the same way the value for r may be due to sampling fluctuations.

If the coefficient of correlation is computed for each of a large number of samples of paired values, a frequency distribution of the resulting coefficients will be normal (if the true association is zero). Through application of the standard deviation (standard error of the coefficient of correlation) it can be foretold that no value of r greater than three times its standard error will arise due to chance (99.7 times out of 100). If, therefore, the computed r is more than three times σ_r, 99.7 times out of 100 it is significant

To determine the error likely to arise due to sampling, the standard error of the coefficient of correlation may be used in the same fashion as the standard error of the mean.

Fifty chances out of 100 the difference between the observed and the actual r will not be larger than .6745 σ_r (the probable error of r), and 99.7 chances out of 100 this difference will not be larger than 3 σ_r.

The formula for the standard error of the coefficient of correlation is:

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$$

However, when the coefficient of correlation for the underlying population approaches 100%, the sampling distribution cannot be normal or symmetrical, since the possibilities of extremes in one direction are limited by the maximum obtainable value for r of 1.00, while the range of possible values of r in the opposite direction is still great. Here the value of r may be converted for tests of reliability and significance into a more useful value of z.

$$z = \tfrac{1}{2}\left[(\log_e (1 + r) - \log_e (1 - r)\right]$$

The sampling distribution of this value approaches normality and is symmetrical.

Its standard error is[1] $\sigma_z = \frac{1}{\sqrt{N - 3}}$

Small Samples—Standard Error of Mean

Due to serious errors which arise in this technique, where the number of items constituting the sample is small (generally less than 30) the standard errors outlined above can no longer be used.

If the sample is small, a new standard error is computed:[2]

$$s^2 = \frac{\Sigma (x^2)}{N - 1} = \frac{N \sigma^2}{N - 1}$$

$$S_{\bar{x}} = \frac{s}{\sqrt{N}}$$

But, for small samples the usual values of the multiples of standard errors taken to include a given percent of the cases can no longer be applied. The multiples to be used for various percentages of probability of occurrence of deviation not greater than given size are shown below.[3] (The N' in the table for the standard error of the mean is $N - 1$.)

[1] For a more elaborate discussion of the standard error of s see Fisher, R. A., *Statistical Methods for Research Workers*.

[2] If σ is used, rather than s, this formula may be written $S_x = \frac{\sigma}{\sqrt{N - 1}}$.

[3] This table is generally referred to as the "t" table and the value in the body of the table as "t."

N'	50%	95%	99%
1	1.000	12.706	63.657
2	.816	4.303	9.925
3	.765	3.182	5.841
4	.741	2.776	4.604
5	.727	2.571	4.032
6	.718	2.447	3.707
7	.711	2.365	3.499
8	.706	2.306	3.355
9	.703	2.262	3.250
10	.700	2 228	3.169
11	.697	2.201	3.106
12	.695	2.179	3.055
13	.691	2.160	3 012
14	.692	2.145	2.977
15	.691	2 131	2 947
16	.690	2 120	2.921
17	.689	2.110	2 898
18	.688	2.101	2.878
19	.688	2.093	2.861
20	.687	2 086	2.845
21	.686	2.080	2.831
22	.686	2 074	2.819
23	.685	2.069	2 807
24	.685	2 064	2.797
25	.684	2 060	2 787
26	.684	2 056	2 779
27	.684	2 052	2 771
28	683	2.048	2 763
29	683	2 045	2.756
30	683	2 042	2.750

Other Standard Errors for Small Samples:

The standard errors for various other statistical measures when computed for small samples are shown below. The results obtained from these formulae may be applied in the same manner as the standard errors for large samples (See pp. 113 to 122), using the appropriate multiples of the standard error obtained from the table above.

Measure	Standard Error[1] (For small samples)	Value of N' in table of multiples
Difference between[2] two means	$s^2 = \frac{\Sigma(x^2_1) + \Sigma(x^2_2)}{(N_1 - 1) + (N_2 - 1)}$ $S_D = \frac{s}{\sqrt{\frac{N_1 N_2}{N_1 + N_2}}}$	$N' = N_1 + N_2$
Coefficient of[3] Correlation	$S_r = \frac{1 - r^2}{\sqrt{N - 2}}$	$N' = N - 2$
Coefficient of Correlation in terms of "z"	$\sigma_z = \frac{1}{\sqrt{N - 3}}$	Use multiples as for large samples.

[1] Fisher, R. A., *Statistical Methods for Research Workers.*

[2] x_1 is the deviation of the actual values from the mean of all X_1 values.

[3] The method involving transposition of r to z is generally more satisfactory than the one given here.

ADDITIONAL BIBLIOGRAPHY*

BOWLEY, ARTHUR L., *Elements of Statistics*, pp. 178–195; 312–342. P. S. King & Son, London, 1907.

FISHER, R. A., *Statistical Methods for Research Workers*, pp. 53–64; 70–78; 106–119; 123–133. Oliver & Boyd, Edinburgh 1932

KELLEY, TRUMAN L., *Interpretation of Educational Measurements*, pp. 54–61; 156–158; 171–178; 188. World Book Co., Yonkers, New York, & Chicago, Illinois, 1927.

ODELL, C. W., *Educational Statistics*, pp. 221–241. Century Co., New York, 1925.

OTIS, ARTHUR S., *Statistical Method in Educational Measurements*, pp. 247–266. World Book Co., Yonkers, New York, & Chicago, Illinois, 1926.

RIETZ, H. L. (Editor), *Handbook of Mathematical Statistics*, pp. 71–77. Houghton Mifflin Co., New York, 1924.

RUCH, G. M., & STODDARD, GEORGE R., *Tests and Measurements in High School Instruction*, pp. 363–374. World Book Co., Yonkers, New York, & Chicago, Illinois, 1927.

TRAUBE, MARION R., *Measuring Results in Education*, pp. 456–465. American Book Co., New York, 1924.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents

CHAPTER XIV

INDEX NUMBERS

Definition

The **index number** is a statistical device for measuring changes in groups of data.

The method may be applied to many general conditions such as employment, prices, group health, academic grades, etc. Data descriptive of these general conditions fluctuate widely; but such data exhibit nevertheless, definite and measurable general tendencies.

In order to measure the changes in the large number of constantly varying items in the data it is necessary to resort to some relative[1] averaging device that will serve as a yardstick of comparative measurement. The index number is such a device.

The index number measures fluctuations during intervals of time, group differences of geographical position or degree, etc. Thus it is possible to obtain an index number showing the relative sales possibilities for a given product in different territories; the academic standing of a group of college students as compared with other groups of students; or to ascertain the relative credit position of a single corporation as compared with many others in the same industry.

For purposes of explanation the discussion below is largely confined to index numbers of prices.

Index Number Construction Problems

1. The purpose for which the index is used has a definite bearing upon the choice of the data used, the method followed, etc.

2. Careful selection of the number and types of items to be used is necessary so that the index fluctuations will be truly representative of the fluctuations in the series.

3. After determination of the proper method of data collection it is necessary to find the available sources of the data needed. Then follows the necessity of actually collecting the data.

4. The problems of selecting the base period and the best method of computation must be solved.

5. The degree of relative importance of each constituent item to the purpose of the index must be determined. This designation of the relative importance of each item is known as *weighting*.

[1] Index numbers are not always relative (percentage) in form. Occasionally they are expressed in terms of absolute (actual) values

Number and Kinds of Commodities

Index numbers must be constructed from samples or limited portions of the types of prices considered; therefore a number of rules should be observed in selecting the commodities:

1. The sample used should be representative (see p. 113). The items selected should be chosen for their representative quality rather than because of the ease of securing quotations.
2. A sufficient number of items should be used. Dr. Irving Fisher points out that index numbers of prices are seldom of much value "unless they consist of more than 20 commodities and 50 is a much better number." He also shows that "after 50 the improvement obtained from increasing the number of commodities is gradual and it is doubtful if the gain from increasing the number beyond 200 is ordinarily worth the extra trouble and expense."[1]

The Base Period

The base period is assigned the value of 100% and is thereby arbitrarily established as a reference period. Index numbers are then computed as relative to the base period.

In the selection of the base period the following should be considered:

1. The base period should not be too far in the past; this in order that a comparison of the price level as relative to the base period will be of definite present or comparative value.
2. Comparison is generally made to a "normal" period; therefore the base period should not be extreme.

Shifting the Base

For comparative purposes the base of an index number series is sometimes shifted from one period to another. The shift may be accomplished by dividing each number in the series by the index number indicated for the period to be used as the new base year, the result is then multiplied by 100.[2]

In the following illustration the base year of the index series, 1926, is shifted to 1928 by dividing each index number by the index value for 1928 (150.0) and multiplying by 100.

1926	1927	1928	1929
100.0	110.1	150.0	125.3
66.7	73.4	100.0	83.5

[1] Fisher, Irving—*The Making of Index Numbers*, p. 340.

[2] All types of index numbers cannot have their base shifted in this manner (see discussion below for types which can be handled in this manner). In order to use a new base period certain types must be completely reconstructed.

Selection of Method of Computation

Irving Fisher gives over 150 different formulae for the construction of index numbers.[1] These formulae, however, are largely variations of a limited number of main types.

Some of the major groups of methods of constructing index numbers may be classified as:

1. The Unweighted (Simple) Method:
   a. The aggregate of actual prices.
   b. The average of relative prices.
2. The Weighted Method:
   a. The weighted aggregate of actual prices.
   b. The weighted average of relative prices.

Simple Aggregate of Actual Prices

The index number constructed by the simple aggregate method is a comparison of the sum of the prices for the commodities considered to the sum of the prices for the same commodities in the base period.[2]

$$\frac{\Sigma p_n}{\Sigma p_o}$$ (Index number formula No. 1)

where

Σp_n = sum of prices of commodities of any given period.

Σp_o = sum of prices of commodities in base period.

Table 33—Computation of Index of Wholesale Metal Prices By Unweighted Aggregate of Actuals Method

(1926 Used as Base Year)

Metal	Unit	Prices In Dollars		
		1926	1928	1930
Pig Iron	Ton	$20.4200	$17.6800	$17.1700
Copper	Pound	.1393	.1468	.1311
Aluminum	Pound	.2699	.2390	.2339
Lead	Pound	.0825	.0614	.0538
Zinc	Pound	.0737	.0603	.0456
Tin	Pound	.6536	.5039	.3163
Silver	Ounce	.6211	.5818	.3815
Total		$22.2601	$19.2732	$18 3322
Index		100.0%	86.6%	82.4%

However, the index number computed by the simple aggregate method is subject to a serious defect in that those commodities which have large figure quotations will dominate the index. For

[1] Fisher, Irving—*The Making of Index Numbers.*

[2] The index may be expressed as a sum of money rather than a relative or percentage figure, i.e., the Bradstreet index of wholesale prices for January, 1931 was $9.51.

instance, if there should be a *decrease* of 10% in the price of pig iron while all other commodities *rose* 10%, thus indicating an increase in the price level, nevertheless the predominating influence of the pig iron quotation will cause the index to fall.

	Number of Commodities	1926	19—	
Pig Iron	1	$20.42	$18.378	Decrease of 10%
All other commodities	6	1.8401	2.0241	Increase of 10%
Total	7	$22.2601	$20.4021	
Index		100%	91.7%	

The difficulty indicated above cannot be avoided by reducing all commodities to a common unit such as a pound—as done in the Bradstreet index. Such procedure would only give rise to new inequalities. Applied to the problem in table 33 pig iron would be $.0102 a pound while tin would be $.6536 a pound (1926).

Average of Relative Prices

A method which avoids the price inequalities shown above involves the conversion of the price figures into relatives. A price relative is a statement of the price of a commodity as a percent of its price in the base period. Expressed in formula form:

$$\frac{p_n}{p_o}$$

where

p_n = price in the given period.
p_o = price in the base period.

The relative for each commodity in the base period is 100%. The relative for the period under consideration is averaged to

Table 34—Computation of Index of Wholesale Metal Prices by Unweighted Arithmetic "Mean of Relatives" Method

(1926 Used as Base Year)

Metal	Unit	1926		1928		1930	
		Price in Dollars	Relative Price	Price in Dollars	Relative Price	Price in Dollars	Relative Price
Pig Iron	Ton	$20 4200	100%	$17 6800	86 6%	$17 1700	84.1%
Copper	Lb.	.1393	100	.1468	105 4	.1311	94 1
Aluminum	Lb.	.2699	100	.2390	88 6	2339	86.7
Lead	Lb.	.0825	100	0614	74.4	0538	65 2
Zinc	Lb.	.0737	100	0603	81 8	0456	61.9
Tin	Lb.	.6536	100	.5039	77 1	.3163	48.4
Silver	Oz.	.6211	100	5818	93 7	.3815	61 4
Totals			700%		607 6%		501.8%
Index			100%		86.8%		71.7%

obtain the index number. The arithmetic mean, the median or the geometric mean may be used for averaging.

The formula below demonstrates the computation of an index number by the average of relative prices method, using the arithmetic mean.

Formula:

$$\frac{\Sigma\left(\frac{p_n}{p_o}\right)}{N}$$

(Index number formula No. 2)

Advantages and Disadvantages of Various Averages In Index Number Construction

Arithmetic Mean

Advantages

1. The mean is relatively easy to compute.

2. Due to long and common usage the arithmetic mean is commonly understood.

3. If a weighted average is taken the means of subgroups can be averaged to obtain the means of values. (A weighted average may be necessary if there are a varying number of items in the various groups).

Disadvantages

1. The mean is greatly affected by extremes (compare p. 21).

2. Increases are given a greater emphasis than decreases. For instance, if commodity A should rise from \$1 to \$2, an increase of 100 percent, while commodity B fell from \$2 to \$1, a decrease of 50 percent, an index of the price level of these two commodities (if the arithmetic mean of relatives is used) will show an increased instead of an unchanged price level.

	1926		1928	
Commodity	Price	Relative	Price	Relative
A	\$1	100%	\$2	200%
B	\$2	100	\$1	50
Total		200%		250%
Index		100%		125%

3. The base of the index number computed by the average of relatives method cannot be shifted by the short method.

Median

Advantages

1. Unlike the arithmetic mean the median will not overemphasize increases.

2. The median is less affected by extremes than the arithmetic mean (see p. 21).

3. It is easy to compute. The relatives are arranged according to size and the middle one is selected as the median.

Disadvantages

1. The median cannot be treated algebraically; i.e. the medians of subgroups cannot be averaged to obtain the median of all the data.

2. Its value is erratic when the number of items is small.

3. The index constructed by this method cannot be shifted to a new base by the short method.

Geometric Mean[1]

Advantages

1. The geometric mean does not overemphasize increases; rather, it gives equal importance to equal ratios of change.

In the problem illustrating the disadvantage of the arithmetic mean (p. 133) the correct index number can be secured by use of the geometric mean.

	1926		1928	
Commodity	Price	Relative	Price	Relative
A	$1	100%	$2	200%
B	$2	100	$1	50
Geometric Mean		100%		100%

2. The base of an index number constructed by this technique can be shifted by the short method.

Disadvantages

1. The calculation of the geometric mean is laborious.

2. It is an unfamiliar form of average.

The Weighting of Index Numbers

It is often desirable to assign a varying degree of importance to the items composing the index numbers. If this action is not taken each commodity will be given a weight or importance depending upon the size of the price, or upon some other chance factor, rather than a proportionate weight depending upon its importance.

An objection to the *unweighted* aggregate of actual prices method of constructing an index number may be eliminated by introducing a *deliberate* system of weights. To measure the weight or importance of the items composing a price index the quantity of each commodity produced may be used.

The Weighted Average

A weighted average (arithmetic mean) may be obtained:

1. By multiplying each item by its corresponding weight.
2. By totaling the results obtained.
3. By dividing by the sum of the weights.

$$\text{Weighted Average} = \frac{\Sigma\,(\text{Items} \times \text{Weights})}{\Sigma\,(\text{Weights})}$$

[1] For a more complete explanation of the geometric mean see p. 26.

Thus if it is found that in a particular section there are two quotations on the price of bread, say 6¢ in chain stores selling 10,000 loaves and 8¢ in independent bakers selling 1000 loaves, a weighted average of the prices may be determined as follows:

	Price	Quantity Sold	Price Times Quantity
Chain Store	$ .06	10,000	600
Bakery	.08	1,000	80
		11,000	680

680 ÷ 11,000 = $ .062

Weighted Aggregate of Actual Prices

A weighted aggregate[1] of actual prices may now be computed by using as a weight the quantity of each commodity produced. The quantities produced in some fixed period, such as the base year, may be used as weights. The index is obtained by comparing the weighted aggregate (total) for the given year to that for the base year.

When $\frac{\Sigma\ (p_n q_o)}{\Sigma\ (p_o q_o)}$ (Index number formula No. 3a)

p_n = Price given year
p_o = Price base year
q_o = Quantity base year
q_n = Quantity given year

For 1928 $\frac{\Sigma\ (p_1 q_o)}{\Sigma\ (p_o q_o)} = \frac{\$1,272,012.51}{\$1,446,076\ 73} = 88.0\%$

For 1930 $\frac{\Sigma\ (p_2 q_o)}{\Sigma\ (p_o q_o)} = \frac{\$1,149,875.80}{\$1,446,076.73} = 79.5\%$

However, since conditions change, the quantity of the commodities produced in any one fixed period will not be a good measure of their relative importance for all other periods. To meet this objection a set of weights which change every year may be used. Thus the quantity produced in each given year may be used as weights when constructing the index for *that particular period.* The formula will then read:

$\frac{\Sigma\ (p_n q_n)}{\Sigma\ (p_o q_n)}$ (Index number formula No. 3b)

For 1928

$\frac{\Sigma\ (p_1 q_1)}{\Sigma\ (p_o q_1)} = \frac{\$1,268,414.03}{\$1,438,339.20} = 88.19\%$

For 1930

$\frac{\Sigma\ (p_2 q_2)}{\Sigma\ (p_o q_2)} = \frac{\$962,303.20}{\$1,220,635.05} = 78.84\%$

[1] The *weighted average* (arithmetic mean) is obtained by dividing the sum of the various items multiplied by their respective weights by the sum of the weights. The *weighted aggregate* or sum is obtained by securing the sum of the various items times their corresponding weights *without* dividing by the sum of the weights.

Table 35—Computation of Index of Wholesale Prices of Metals in the United States by Weighted Aggregate of Actuals Method Using Base Year Weights

(1926 Used As Base Year)

Metal	Unit	1926			1928		1930	
		Price In Dollars p_0	Production (Thousands) q_0	Price Times Quantity $p_0 q_0$	Price In Dollars p_1	Price Times Quantity $p_1 q_0$	Price In Dollars p_2	Price Times Quantity $p_2 q_0$
Pig Iron	Ton	$20.4200	39,373	$803,996.66	$17.6800	$696,114.64	$17.1700	$676,034.41
Copper	Pound	.1393	1,744,860	243,059.00	.1468	256,145.45	.1311	228,751.15
Aluminum	Pound	.2699	145,000	39,135.50	.2390	34,655.00	.2339	33,915.50
Lead	Pound	.0825	1,416,280	116,843.10	.0614	86,959.59	.0538	76,195.86
Zinc	Pound	.0737	1,236,800	91,152.16	.0603	74,579.04	.0456	56,398.08
Tin*	Pound	.6536	172,790	112,935.54	.5039	87,068.88	.3163	54,653.48
Silver	Ounce	.6211	62,719	38,954.77	.5818	36,489.91	.3815	23,927.30
Total				$1,446,076.73		$1,272,012.51		$1,149,875.78

* Imports.

Table 36—Computation of Index of Wholesale Prices of Metals in the United States by Weighted Aggregate of Actual Method Using Given Year Weights

(1926 Used As Base Year)

Metal	Unit	1926	1928				1930			
		Price in Dollars p_0	Price in Dollars p_1	Production (Thousands) q_1	$p_1 q_1$	$p_0 q_1$	Price in Dollars p_2	Production (Thousands) q_2	$p_2 q_2$	$p_0 q_2$
Pig Iron. . .	Ton	$20 4200	$17.6800	38,156	$674,598 08	$779,145 52	$17.1700	31,399	$539,120.83	$641,167.58
Copper.	Pound	.1393	.1468	1,818,280	266,923.50	253,286.40	.1311	1,380,960	181,043.86	192,367.73
Aluminum. . .	Pound	.2699	.2390	210,000	50,190.00	56,679.00	.2339	229,000	53,563.10	61,807.10
Lead.	Pound	.0825	.0614	1,302,280	79,959.99	107,438.10	.0538	1,230,220	66,185.84	101,493.15
Zinc.	Pound	.0737	0603	1,239,180	74,722.55	91,327.57	.0456	1,008,640	45,993.98	74,336.77
Tin*. . . .	Pound	.6526	5039	174,650	88,006.14	114,151.24	.3163	180,940	57,231.32	118,262.38
Silver**. . . .	Ounce	.6211	.5818	58,463	34,013.77	36,311.37	.3815	50,234	19,164.27	31,200.34
Total. . . .					$1,268,414.03	$1,438,339.20			$962,303.20	$1,220,635.05

* Imports

** United States Production only.

Table 37—Computation of Index of Wholesale Prices of Metals in the United States by Weighted Average of Relative Methods Using Given Year Weights

(1926 Used as Base Year)

Metal	Unit	1926		1928					1930				
		Price in Dollars	Price Relative	Price in Dollars	Price Relative	Production (Thousands)		Relative Times Weight	Price in Dollars	Price Relative	Production (Thousands)		Relative Times Weight
		p	$\frac{p_0}{p_0}$	p_1	$\frac{p_1}{p_0}$	q_1	p_1q_1	$\frac{p_1}{p_0} \times p_1q_1$	p_2	$\frac{p_2}{p_0}$	q_2	p_2q_2	$\frac{p_2}{p_0} \times p_2q_2$
Pig Iron	Ton	$20.4200	100%	$17.6880	86.6%	38,156	674,598.08	584,201.94	$17.1700	84.1%	31,399	539,120.83	453,400.62
Copper	Pound	.1393	100	.1468	105 4	1,818,280	266,923.50	281,337.37	.1311	94.1	1,380,960	181,043.86	170,362 27
Aluminum	Pound	.2699	100	.2390	88 6	210,000	50,190.00	44,468.34	.2399	86.7	229,000	54,937.10	47,630.47
Lead	Pound	.0825	100	.0614	74.4	1,203,280	79,959.99	59,490.23	.0538	65.2	1,230,220	66,185.84	43,153.17
Zinc	Pound	.0737	100	.0603	81.8	1,239,160	74,722.55	61,123.05	.0456	61.9	1,008,640	45,993.98	28,470.27
Tin	Pound	.6536	100	.5039	77.1	174,650	88,006.14	67,852.73	.3162	48.4	180,940	57,231.32	27,699.96
Silver	Ounce	.6211	100	.5818	93.7	58,463	49,722.37	46,589.86	.3815	61.4	50,234	19,164.27	11,766.86
Total							1,284,122.63	1,145,063.52				963,677.20	782,483.62

Weighted Average of Relatives

An index number may be constructed by securing a weighted average of the relative prices for the period under consideration. Quantities of production, however, can no longer be used as weights since each quantity is expressed in different units (tons, pounds, ounces, bushels, etc.). The column of figures resulting from the multiplication of the price relatives by these weights would be expressed in different units and could not be totaled. It becomes necessary to use weights expressed in common units. The most usual common unit is the dollar. The money *value* rather than the *quantity* of production may then be used as weights.

With base period weights and using a weighted arithmetic mean the formula will be:

$$\frac{\Sigma\left[\frac{p_n}{p_o} \times (p_o q_o)\right]}{\Sigma\,(p_o q_o)} \quad \text{(Index number formula No. 4a)}$$

for $p_o q_o$ equals value of production in the base period (the price times the quantity).

Through cancellation the formula reduces to:

$$\frac{\Sigma(p_n q_o)}{\Sigma(p_o q_o)}$$

or the same as the weighted aggregate using base year weights (formula 3a).

If given year weights are used a new formula is evolved.

$$\frac{\Sigma\left[\frac{p_n}{p_o} \times (p_n q_n)\right]}{\Sigma(p_n q_n)} \quad \text{(Index number formula No. 4b)}$$

For 1928

$$\frac{\Sigma\left[\frac{p_1}{p_o} \times p_1 q_1\right]}{\Sigma(p_1 q_1)} = \frac{\$1,145,063.52}{\$1,284,122.63} = 89.17\%$$

For 1930

$$\frac{\Sigma\left[\frac{p_2}{p_o} \times p_2 q_2\right]}{\Sigma(p_2 q_2)} = \frac{\$782,483.62}{\$963,677.20} = 81.20\%$$

The Ideal Index Number

Irving Fisher has developed an index number which meets the requirements of certain tests (see page 140) which can be applied to index numbers. His formula is a "cross" or geometric average of two formulae which are subject to opposite error. It is the geometric average of the aggregate of actuals weighted by base

year quantities (formula 3a) and the aggregate with given year weights. The formula is:

$$\sqrt{\text{Formula 3a} \times \text{Formula 3b}}$$

$$\sqrt{\frac{\Sigma (p_n q_o)}{\Sigma (p_o q_o)} \times \frac{\Sigma (p_n q_n)}{\Sigma (p_o q_n)}} \qquad \text{(Index number formula No. 5)}$$

Index Number Tests

Time Reversal Test

If computed by using the given period as a base the result obtained from an index number using a certain period as a base should not give inconsistent results. For example, if an index number with a base at 1926 (1926 = 100) should give rise to an index of 2.00 for 1928, reconstructing the index with a base of 1928 the index for 1926 should be .50, the reciprocal of 2.00.

	1926		1928
Index A	1.00	⟍ ⟋	2.00
B	.50	⟋ ⟍	1.00

Cross multiplying the index numbers as indicated by the arrows should give a value of 1.00, since these numbers are reciprocal.[1]

Factor Reversal Test

The change in the price times the change in the quantity should be equal to the change in the value of the commodities produced. The index of prices can be obtained by any of the methods; for example, the aggregate of actuals weighted by base or fixed year weights (formula 3a) may be used for the purpose.

$$\frac{\Sigma (p_n q_o)}{\Sigma (p_o q_o)}$$

An index of the quantity of production can be secured by reversing the positions of the price figures (p) with the quantity figures (q).

$$\frac{\Sigma (q_n p_o)}{\Sigma (q_o p_o)}$$

An index of the value of production can be obtained by comparing the value of production in the given period (V_n) to the value of production in the base period (V_o).

Therefore:

$$\frac{\Sigma (p_n q_o)}{\Sigma (p_o q_o)} \times \frac{\Sigma (q_n p_o)}{\Sigma (q_o p_o)} \text{ should equal } \frac{V_n}{V_o}$$

where V_n the value of production, in the given period, may be obtained by multiplying the price in the base period by the quan-

[1] A reciprocal of a number is 1 divided by that number.

tity produced or $\Sigma(p_n q_n)$ and for the base period $\Sigma(p_o q_o)$.

The test will then read:

$$\frac{\Sigma(p_n q_o)}{\Sigma(p_o q_o)} \times \frac{\Sigma(q_n p_o)}{\Sigma(q_o p_o)} \text{ should equal } \frac{\Sigma(p_n q_n)}{\Sigma(p_o q_o)}$$

Table 38 lists some of the more important price index number series.

Quantity Index Numbers

Index number technique can be applied to measurement of changes in quantity groups as well as price changes. Index numbers of this type are used to measure changes in business activity, industrial production, commodity stocks, etc.

The methods of construction are the same for quantity index numbers as for price index numbers. The simplest form is the simple aggregate type

$$\frac{\Sigma q_n}{\Sigma q_o}$$

where Σq_n is the sum of the quantities in any given period

Σq_o is the sum of the quantities in the base period.

Since this form of index involves the sums of series the various quantities must all be in the same units (tons, bushels, etc.) to make the summation possible.

When the units are different for the various items in the series and an unweighted index number is desired the average of relatives may be used. If the arithmetic mean is used as the average the formula is

$$\Sigma \left(\frac{q_n}{\quad}\right)$$

$$N$$

It is generally desirable however, to weight the index numbers in order to arbitrarily assign various degrees of importance to the several items composing the index number. Either the price of the commodity or some arbitrary weight may be used for this purpose.

The weighted aggregate form for use in measuring quantity changes is

$$\frac{\Sigma(q_n p_o)}{\Sigma(q_o p_o)}$$

with base year weights

or

$$\frac{\Sigma(q_n p_n)}{\Sigma(q_o p_n)}$$

with given year weights

Table 38—Price Indexes

Name of Index	Periodicy	Base	Number of Commodities	Area Covered	Compiling Agency	Where Published
Wholesale						
Bureau of Labor Statistics All Commodity Index	Monthly	1926 = 100	784	U. S.	U. S. Bureau of Labor Statistics	Survey of Current Business
Bradstreet's Commodity Price Index	Monthly	.....	96	U. S.	Dun and Bradstreet Inc.	Survey of Current Business
Dun's Commodity Price Index	Monthly	.....	300	U. S.	Dun and Bradstreet Inc.	Dun and Bradstreet Monthly Review
Dun and Bradstreet Price Index	Monthly	.....	13 groups	U. S.	Dun and Bradstreet Inc.	Dun and Bradstreet Monthly Review
World Prices Index	Monthly	1923–25 = 100	9	World	U. S. Department of Commerce	Survey of Current Business
Retail						
Bureau of Labor Statistics Food Price Index	Bi-Monthly	1913 = 100	48	U. S. and Hawaii	U. S. Bureau of Labor Statistics	Survey of Current Business
Fairchild's Retail Price Index	Monthly	Dec. 1930 = 100	5 groups	U. S.	Fairchild Publications	Survey of Current Business
Cost of Living						
National Industrial Conference Board Cost of Living Index	Monthly	1923 = 100	5 groups	U. S.	National Industrial Conference Board	Survey of Current Business
U.S. Bureau of Labor Index Cost of Living	Semi-Annually	1913 = 100	6 groups	U. S.	U. S. Bureau of Labor	Survey of Current Business
Farm Prices						
Index of Prices Received by Farmer	Monthly	Aug. 1909–July 1914 = 100	8 groups	U. S.	U. S. Department of Agriculture	Crops and Markets
Index of Prices Paid for Commodities Used	Monthly	1910–1914 = 100	3 groups	U. S.	U. S. Department of Agriculture	Crops and Markets
Purchasing Power of the Dollar Based on						
Wholesale Prices	Monthly	1923–25 = 100	784	U. S.	U. S. Department of Labor	Survey of Current Business
Retail Prices	Monthly	1923–25 = 100	4 groups	U. S.	U. S. Department of Labor	Survey of Current Business
Farm Prices	Monthly	1923–25 = 100	8 groups	U. S.	Department of Agriculture	Survey of Current Business
Cost of Living	Monthly	1923–25 = 100	5 groups	U. S.	National Industrial Conference Board	Survey of Current Business

Unless the units are the same for all items only the prices can be used as weights and not arbitrary weights since if the latter are used no summation will be possible.

The weighted average of relatives may be used where the units are different and it is desired to use arbitrary weights

$$\frac{\Sigma\left(\frac{q_n}{q_o} \times \text{wt}\right)}{\Sigma\,(\text{wt})}$$

where

wt weight

The "Ideal Index" can also be converted to quantity form

$$\sqrt{\frac{\Sigma\,(q_n p_o)}{\Sigma\,(q_o p_o)} \times \frac{\Sigma\,(q_n p_n)}{\Sigma\,(q_o p_n)}}$$

The composition of some of the better known quantity index series as well as the system of weighting is shown in table 39. The various items constituting the index have been grouped in this table into major groupings to facilitate comparison.

Table 39—Composition of Selected Indexes of Business Activity and Industrial Production

Component Series	WEIGHTS			
	Indexes of Industrial Production		Indexes of Business Activity	
	Standard Trade and Securities	Federal Reserve Board	New York Times Annalist	Business Week
Iron and Steel Production	25	20.5	25	10
Textile Production	18	17.5	13	
Lumber Production	10	8.5	7	
Agricultural & Food Products	9	9.5		
Automobile & Tire Production	8	5.2	10	
Non Ferrous Metal Production	6	5.2	5	
Paper & Printing Production	4	9.6		
Leather and Shoe Production	4	3.9	2	
Cement Production	4	1.1	3	
Coal Production	3	6.9		3
Railroad Equipment Production	2	0.6		
Electric Power Production	2		15	12
Chemical Production	1	3.2		
Enameled Ware Shipments	1			
Stone, Clay & Glass Production		2.2		
Rubber		1.5		
Carloadings			20	17
Debits to Individual Accounts				30
Building Construction				20
Commercial Loans				4
Money in Circulation				4

ADDITIONAL BIBLIOGRAPHY*

BOWLEY, ARTHUR L., *Elements of Statistics*, pp. 96–213. P. S. King & Son, London, 1901.

FISHER, IRVING, *The Making of Index Numbers.* Houghton, Mifflin Company, Boston, 1927.

MITCHELL, WESLEY C., *Index Numbers of Wholesale Prices in the United States and Foreign Countries.* United States Bureau of Labor Statistics, Bulletin No. 173; Washington, D. C., 1921.

REITZ, H. L. (Editor), *Handbook of Mathematical Statistics*, pp. 181–194. Houghton Mifflin Co., New York, 1924.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER XV

FURTHER ANALYSIS OF THE FREQUENCY DISTRIBUTION

Moments

A frequency distribution can be more accurately analyzed if certain constants or "moments" of the distribution are computed. Moments are used for computing measures which are descriptive of the distribution, and for the determination of the appropriate curve to be used in smoothing the distribution mathematically (see p. 105).

I. The first moment of a frequency distribution as measured about any arbitrary origin is:*

$$\nu_1 = \frac{\Sigma(fd)}{N}$$

II. The second moment (about an arbitrary origin) is:

$$\nu_2 = \frac{\Sigma f(d^2)}{N}$$

III. The third moment (about an arbitrary origin) is:

$$\nu_3 = \frac{\Sigma f(d^3)}{N}$$

IV. The fourth moment (about an arbitrary origin) is

$$\nu_4 = \frac{\Sigma f(d^4)}{N}$$

The most important moments are those which are measured with the mean as the origin:**

$$\mu_1 = \frac{\Sigma f(x)}{N} = 0$$

$$\mu_2 = \frac{\Sigma f(x^2)}{N}$$

$$\mu_3 = \frac{\Sigma f(x^3)}{N}$$

$$\mu_4 = \frac{\Sigma f(x^4)}{N}$$

where x represents the deviation of the actual value from the mean.

* The symbol ν is the Greek small letter Nu

** The symbol μ is the Greek small letter Mu.

The sum of the deviation about the mean is zero and, therefore, the first moment will equal zero. The other moments about the mean can be obtained readily from

$$\mu_2 = \nu_2 - \nu_1^2$$

$$\mu_3 = \nu_3 - 3\ \nu_1\nu_2 + 2\ \nu_1^3$$

$$\mu_4 = \nu_4 - 4\ \nu_1\nu_3 + 6\ \nu_1^2\nu_2 - 3\nu_1^4$$

Sheppard's Corrections for Grouping[1]

The computation of the moments from a frequency distribution involves the assumption that the values may be dealt with as though they were all located at the midpoint of the class interval. This assumption is subject to a certain error, allowance for which can be made by use of the corrections shown below:

I. Corrected First Moment

$$\mu_1' = 0$$

II. Corrected Second Moment

$$\mu_2' = \mu_2 - 1/12$$

III. Corrected Third Moment

$$\mu_3' = \mu_3$$

IV. Corrected Fourth Moment

$$\mu_4' = \mu_4 - \tfrac{1}{2}\quad \mu_2 + 7/240$$

For convenience, the moments calculated by the methods outlined above are generally computed in terms of class intervals rather than in original units. To convert the moments back to the original units the following relationships are used:

μ_2' (in original units) $= C^2\ \mu_2'$ (in class interval units)

μ_3' (in original units) $= C^3\ \mu_3'$ (in class interval units)

μ_4' (in original units) $= C^4\ \mu_4'$ (in class interval units)

where C = size of class interval groupings.

The computation of the moments is illustrated on the following page.

[1] The corrections apply only when (a) the distribution is continuous (see p. 7), and when (b) the distribution tapers off gradually in both directions.

Table 40—Calculation of Moments

Variation of Thickness in 600 Brass Washers Manufactured by the ABC Co.

(Hypothetical Data*)

Thickness (In Inches)	Number of Washers (f)	Deviation From Arbitrary Origin in Class Intervals d'	$f(d')$	(5) $f(d'^3)$	(6) $f(d'^2)$	(7) $f(d'^4)$
.0180–.0183	6	– 5	– 30	150	– 750	3750
.0184– 01879	30	– 4	– 120	480	– 1920	7680
.0188–.01919	42	– 3	– 126	378	– 1134	3402
.0192–.01959	66	– 2	– 132	264	– 528	1056
.0196– 01999	94	– 1	– 94	94	– 94	94
.0200–.02039	120	0	0	0	0	0
.0204–.02079	102	1	102	102	102	102
.0208–.02119	60	2	120	240	480	960
.0212–.02159	54	3	162	486	1458	4370
.0216–.02199	14	4	56	224	896	3584
.0220–.02239	12	5	60	300	1500	7500
	600		– 2	2718	10	32502

* Hypothetical data based on a smaller distribution given by W. A. Shewhart, *Economic Control of Quality of Manufactured Product*

$$\nu_1 = \frac{\Sigma(fd')}{N} = \frac{-2}{600} = -.0330$$

$$\nu_2 = \frac{\Sigma f(d'^2)}{N} = \frac{2718}{600} = 4.5300$$

$$\nu_3 = \frac{\Sigma f(d'^3)}{N} = \frac{10}{600} = .0167$$

$$\nu_4 = \frac{\Sigma f(d'^4)}{N} = \frac{32502}{600} = 54.1700$$

$$\mu_1 = 0$$

$$\mu_2 = \nu_2 - \nu_1^2 = 4.5300 - (.0033)^2 = 4.52999$$

$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 = .0167 - 3\,(.0033)\,(4.5300) + 2\,(.0033)^3 = -.43177$$

$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4 = 54.1790 - 4\,(.0033)\,(.0167) + 6\,(.0033)^2\,(4.5300) - 3\,(.0033)^4 = 54.170076$$

$$\mu_1' = 0$$

$$\mu_2' = \mu_2 - 1/12 = 4.52999 - .08333 = 4.44666$$

$$\mu_3' = \mu_3 = -.43177$$

$$\mu_4' = \mu_4 - 1/2\mu_2 + 7/240 = 54.170076 - 2.264995 + .029167 = 51.94429$$

Curve Type Criteria

The curve type that best describes the distribution may be identified from criteria calculated on the basis of the values of the moment.

The criteria may be computed as follows:*

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\beta_2 = \frac{\mu_4}{\mu^2_2} = \frac{\mu_4}{\sigma^4}$$

$$\kappa = 4\,(4\beta_2 - 3\beta_1)\,(2\beta_2 - 3\beta_1 - 6) - \frac{\beta_1\,(\beta_2 + 3)^2}{}$$

By using these criteria the type of Pearson curve best describing the distribution may be identified (see Elderton, W. P., *Frequency Curves and Correlation*).[1]

Kurtosis

The kurtosis of a frequency distribution is its "peakedness." If the curve has a higher degree of kurtosis than the normal curve ($\beta > 3$) the curve may be said to be **leptokurtic.** If β is less than 3, the curve is more flat-topped than the normal curve and is said to be **platykurtic.**

The measure of kurtosis is sometimes given as:[2]

$$\beta_2 - 3$$

Where the result is:

1. zero the curve is mesokurtic.
2. a positive value the curve is leptokurtic.
3. a negative value the curve is platykurtic.

The calculation of the β_2 value for the distribution of table 40 is shown below:

$$\beta_2 = \frac{51.94429}{(4.44666)^2} = 2.63$$

Other measures of Skewness

A more exact determination of skewness can be computed from**

$$\alpha_3 = \frac{\mu_3}{\sigma^3} = \sqrt{\beta_1}$$

[1] See p. 105, section on Generalisation of Curves.

[2] Given by other texts also as: $\frac{\beta_2 - 3}{2}$

* The symbol β is the Greek small letter beta and κ is Greek small letter kappa.

** The symbol α is the Greek small letter alpha.

The value of α_3 will be zero for a normal curve.

Another formula for skewness is[1]

$$\chi = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2\,(5\beta_2 - 6\beta_1 - 9)}$$

This value χ (the measure of skewness) can be used to locate the mode more accurately than the methods outlined previously.

$$\text{Mode} = \overline{X} - (\chi)\,(\sigma)$$

[1] Suggested by Karl Pearson. χ is not to be confused with χ^2 used in testing goodness of fit.

* χ is the Greek small letter chi.

ADDITIONAL BIBLIOGRAPHY*

ELDERTON, W. P., *Frequency Curves and Correlation.* C. and E Layton, London 1929.

FISHER, R. A., *Statistical Methods for Research Workers*, pp. 80–105, 228–262. Oliver & Boyd, Edinburgh, 1932.

KELLEY, TRUMAN L., *Interpretation of Educational Measurements*, p. 77. World Book Co., Yonkers, New York, & Chicago, Illinois, 1927.

REITZ, H. L. (Editor), *Handbook of Mathematical Statistics*, pp. 97–111. Houghton Mifflin Co., New York, 1924.

* For readings in standard Statistics textbooks see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER XVI

COLLECTION OF DATA

Assembling and Collecting Data

Data may be obtained from primary original sources, i.e., by interview, questionnaire or letter; or from secondary sources, i.e., data compiled by other individuals or agencies.

Primary Sources

Interview Method

Advantages

1. A higher degree of accuracy is attained through the acquisition of material direct from the source.
2. Material is often obtained that cannot be secured through the questionnaire.
3. There is opportunity personally to check information acquired.

Disadvantages

1. Only small samples can be gathered.
2. The subjective factor is involved in recording by interview.
3. The method is generally inefficient, and the time and expense involved necessarily mean limited field coverage.

Questionnaire Method

Characteristics

1. The questions should be easily understood.
2. If possible they should be arranged in logical sequence.
3. The answers should consist of yes or no, check or blank space, or numerical indication where possible.
4. The questionnaire should be concise.
5. It should be in the most convenient, answerable form.
6. It should be constructed so as to facilitate the tabulation of data.

Advantages

1. A large area may be easily and quickly covered.
2. The method of assembling data is relatively inexpensive.

Disadvantages

1. Frequently questions cannot be answered without a supplementary explanation.
2. In many cases the results are unreliable.

3. A large part of the sample taken may not answer the questionnaire.

Secondary Sources

Advantages

1. The data are already compiled, thereby saving time and expense.
2. The responsibility for accuracy may be shifted.

Disadvantages

1. The data obtained by the primary agent cannot be verified.
2. The statistical technique used may not be obtainable and therefore the accuracy of the results may not be verifiable.
3. Subjective compiling and interpretation may have influenced the result shown.
4. The purpose of the study may have prejudiced the choice of source material and technique adopted.
5. A representative sample may not have been taken.

ADDITIONAL BIBLIOGRAPHY*

BOWLEY, ARTHUR L., *Elements of Statistics*, pp. 14–51. P. S. King & Son, London, 1901.

ODELL, C. W., *Educational Statistics*, pp. 3–33. Century Co., New York, 1925.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER XVII

STATISTICAL TABLES

Definition

The statistical table is a systematic arrangement of numerical data presented in columns and rows for purpose of comparison.

Statistical tables, classified according to purpose, are of two types, general purpose (primary) tables and special purpose (derived or text) tables.

General Purpose Tables

Functions

1. The primary function of the general purpose table is to present original data in tabular form for reference purposes.
2. It serves as a source of information where original data is needed.
3. It is used in the construction of special purpose tables.

Characteristics

1. The general purpose table presents varied information on the same subject.
2. It should contain absolute, not percentage, figures because of its purpose as outlined above.
3. Information should be presented in such form that it can easily be used for reference.
4. Actual figures, not round numbers, should be included.

Special Purpose Tables

Functions

1. The primary function of the special purpose table is to present data so as to emphasize specific relationships.
2. It is used to emphasize a particular phase of the general information contained in a general purpose table.
3. It permits the presentation of selected materials in simple form.

Characteristics

1. Round numbers may be used at times.
2. The selected material in a special purpose table is presented in a small space to facilitate interpretation.

Table 41

TITLE → **Pig Iron Production and Prices in the United States, 1919-1930**

BOXHEAD →

Year	Production (Thousands of Gross Tons)	Prices* (Dollars per Gross Ton)
1919	31,015	$28.97
1920	36,926	42.76
1921	16,688	22.58
1922	27,220	24.06
1923	40,361	26.30
1924	31,406	20.90
1925	36,700	20.58
1926	39,378	20.42
1927	36,566	18.55
1928	38,156	17.68
1929	42,614	18.43
1930	31,399	17.17

STUB →

COLUMN CAPTIONS ←

UNITS ←

*Composite of weekly average prices on foundry and basic pig iron at Valley furnace, Chicago, Birmingham. ← FOOTNOTE

SOURCE → Source: *Iron Age.*

Rules for Table Construction

Practice varies, but the generally accepted rules for the construction of a statistical table are as follows:[1]

1. **Title**—the title should be self explanatory and should indicate in the following order:
   a. the nature of the data presented
   b. the locality covered
   c. the time period included.

The title is placed above the table. The lettering is usually larger in the title than in any other section.

2. **Source:** The source of the material should always be indicated on a table, except where original data has been obtained, since it is used:
   a. to indicate the authority for the data
   b. as a means of verification
   c. as a reference for additional data.

The source is placed below the table at the left.

3. **Footnotes:** A footnote is used to further explain a figure in the table etc. It is placed immediately beneath the table, above the source. The footnote should be indicated by symbols as *, # etc , or by a letter of the alphabet, never by a number, since the latter might be interpreted as part of the table.

[1] Exceptions to the generally accepted procedure of table construction are usually justified by the particular purpose of a specific table.

4. **Arrangement of data:** Items in a table if arranged carefully facilitate reading of the table, analysis and comparison of data and permit emphasizing of selected groups of data. Items may be arranged:

a. **alphabetically**—according to the alphabetic order of the items. This is the most frequently used arrangement for general purpose tables.

b. **chronologically**—according to time of occurrence in comparing subjects over a period of time. Dates should move from the earliest to the latest date from the top of the stub to the bottom or in the boxhead from the left to the right of the table.[1]

c. **geographically**—according to location in the customary classification for example: country, state, county, etc., or Maine, New Hampshire, Vermont, etc. This arrangement is generally confined to reference (general purpose) tables.

d. **by magnitude**—according to size. The largest number is placed at the top of the column and the others arranged in order of size. The row captions correspond to their values. When the row captions are numerical, as class intervals in the frequency distribution, they are arranged by size. The smallest number is arranged at the top for the rows with the largest at the bottom; for the columns, the smallest is placed at the left to the largest at the right.

e. **by customary classification**—There is a customary arrangement for many types of data which do not follow any serial arrangement. For instance the classification men, women, and children, is rarely listed in the order women, children, and men.

5. **Columns:** When there are a number of columns in a table they may be numbered or lettered for reference purposes.

6. **Column Captions:** The heading of each column is known as the column caption. It should be concise. A miscellaneous column is placed at the right end of the table.

7. **Stub:** The heading of a row is known as the row caption. The section of the table containing row headings is designated as the stub. Items in the stub should be grouped, as months grouped by quarters, to facilitate interpretation of the data.

8. **Totals:** The totals of columns should be placed at the bottom of the columns, while row totals should be placed at the extreme right.[2]

[1] An exception to this rule occurs when the latest figures are of primary interest as when the figures are published for the first time In this instance the latest figure may be listed before the others, and then separated from them by a double or heavy line

[2] The United States Census Bureau places totals at column headings and on the extreme left. This practice is explained by the Department as due to the major interest in totals (see footnote 1 above).

9. **Units of Measurement:** These should be included in the boxhead under the column captions.

10. **Rulings:** Lines should be ruled on a table as follows:
    a. A horizontal line is placed below the title, and below the body of the table.
    b. Columns are separated by single lines. In typewritten columns these lines are not essential but are useful.
    c. The stub and boxhead are separated from the figures by double or heavy lines, especially in non printed tables.
    d. Totals are separated from the other figures in a column by a single line.

11. **Emphasis:** A double line, heavy line, *italics* and light and bold face type contrasts are all used for emphasis on tables.

ADDITIONAL BIBLIOGRAPHY*

BOWLEY, ARTHUR L., *Elements of Statistics*, pp. 52–81; 117–124. P. S. King & Son, London, 1901.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER XVIII

GRAPHIC PRESENTATION

Definition

A graph is a method of presenting statistical data in visual form.

Types of Graphs

There are many varieties of graphs. The use of a particular type is dependent upon the data and upon the purpose for which the graph is constructed.

Graphs may be divided into the following types

I. **Line** or **Curve Graphs**
   a. Arithmetic ruling
   b. Semi-logarithmic or logarithmic ruling
   c. Other special rulings

   Special Types of Line Graphs
   a. Silhouette chart
   b. Band chart
   c. High Low Chart
   d. Histogram

II. **Bar charts**

III. **Area Diagrams**

IV. **Solid Diagrams**

V. **Statistical Maps**

Rules for Construction of Graphs[1]

1. Every graph must have a clear and concise title which is generally placed at the top center of the graph.[2] As a rule the title includes information as to:

   a. The nature of the data.
   b. The geographical location
   c. The period covered

These elements of the title customarily appear in the order given above.

[1] For a more complete discussion of the technique of graphic presentation the reader is referred to Arkin, H., and Colton R., *Graphs*, Harper & Bros., 1935

[2] Graphs in printed form generally have the title placed below the graph see graphs in this text

2. **Coordinate lines** should be held to a minimum and **curve lines** emphasized so that the curves stand out sharply against the background.

3. The **source** of the data should be indicated just under the graph at the lower left.

4. **Footnotes,** if any, should be placed under and to the right of the graph.

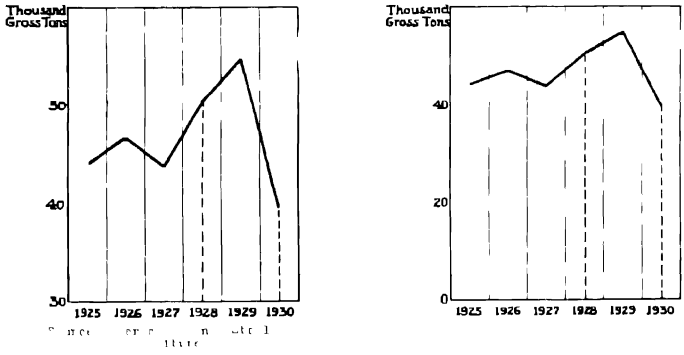
5. If the graph is to be readily understood the curve lines, segments and other details should be as few in number as possible.

6. Each scale must have a **scale caption** indicating the units used.

   a. The *X* axis scale caption should be centered directly beneath the *X* axis.

   b. The *Y* axis scale caption should be placed at the top of the *Y* axis.

7. The **zero point** should be indicated on the scale (*Y* axis), otherwise a misleading comparison may result. The necessity of indicating the zero point is seen by comparison of the peaks at *a* and *b* in the two graphs below, figure 28.[1]



Graph 1
(No Zero line shown)

Graph 2
(Zero line indicated)

Fig. 28—Steel Ingot Production in the United States, 1926—1931.

Inclusion of the zero point (*Y* axis) in graph 2 indicates an entirely different ratio in the heights of the points at 1928 and 1930 (5 to 4) instead of the misleading comparison 2 to 1 in graph 1.

[1] An exception to this rule occurs when the graph is in percentage form. In this case the 100 % line is emphasized

If, however, lack of space makes it inconvenient to use the zero point line a **scale break** may be inserted to indicate its omission. Various types of scale breaks are shown in figure 29.

9. The scales of values should be placed along the X and Y axis, thus giving a general indication of the size of the variations occurring in the graph.

(It is unnecessary to indicate fine gradations on the scales of value since it is not intended that actual values be read off from the graph. Actual values can be obtained from the table of original data which usually accompanies the graph.)

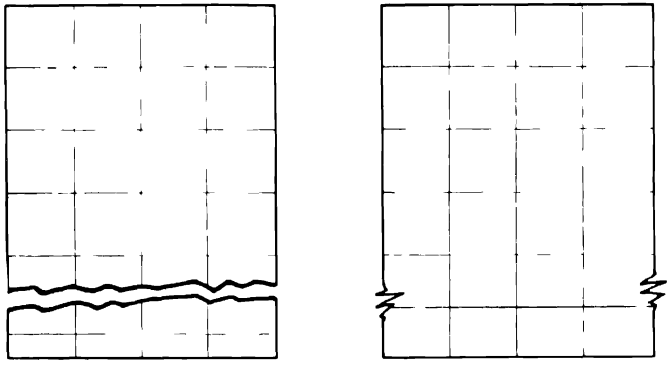


Fig. 29—Types of Scale Breaks.

10. If a space on the X axis is used to indicate time intervals, the point representing the value for each period should be plotted at the midpoint for the period. If desired, however, the periods may be made to coincide with, and the points may then be plotted on, given coordinate lines.

11. On the Y axis the scale of values should run from zero (or from the smallest value) on the bottom of the graph to the highest value at the top. On the X axis the values should run from lowest on the left to highest on the right.

The various elements composing the graph are shown in figure 30.

I. Line or Curve Graphs

The line or curve graph is distinguished by the fact that the variations in the data are indicated by means of a line or curve (see figure 30).

This type of graph is constructed by plotting points whose positions are determined by their respective values on the X and Y scales. The points are connected by straight lines.

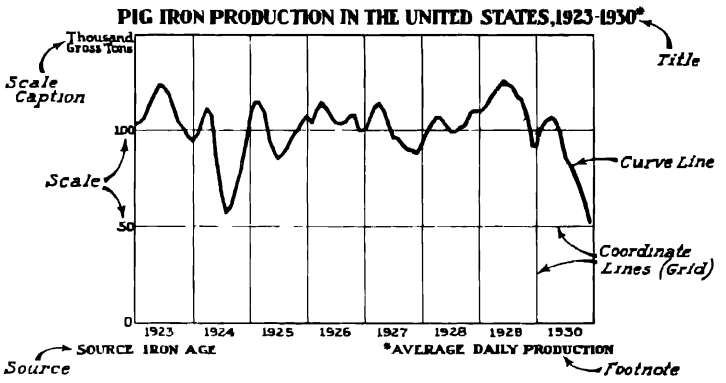


Fig. 30—Elements of the Graph.

Line graphs may be classified according to the type of scale ruling used:

a. Arithmetic ruling

b. Logarithmic rulings

c. Other rulings[1]

Arithmetic Rulings

Arithmetically ruled paper has equal distances between the coordinate lines. Equal quantities will then have equal distances. Thus the distance between 1 and 3 on the background ruling will be the same as that between 8 and 10.

An arithmetic progression will plot as a straight line on arithmetic paper since there are constant differences between the successive values in this type of series.

Since equal amounts are assigned equal distances, equal changes indicate identical absolute differences.

The line or curve type of graph is the most commonly used form of graphic presentation.

Logarithmic and Semi-logarithmic Rulings[2]

When it is desired to compare percentage rather than absolute changes a somewhat different form of ruling is used.

It can be shown[3] that where there is a constant percentage change between two pairs of figures the differences between the logarithms of the figures will be equal.

[1] Various other rulings are also available but are beyond the scope of this discussion.

[2] Semi-logarithmic paper is also known as ratio paper.

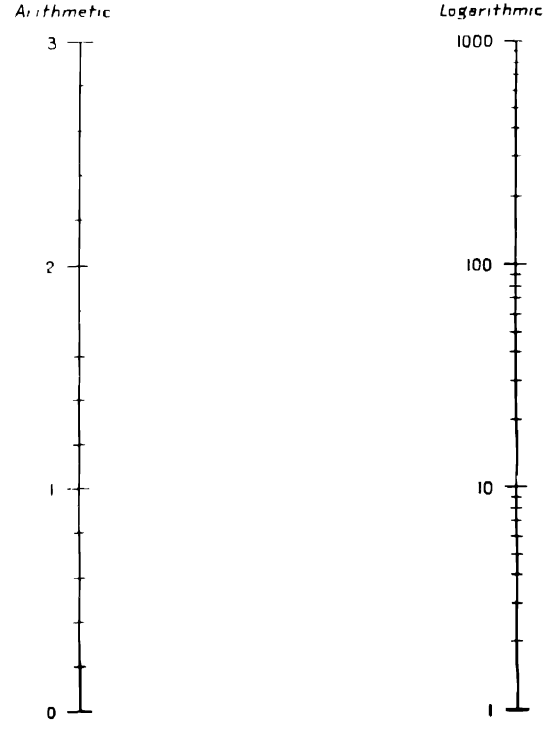
[3] See any text on elementary mathematics

Numbers	Logarithms	
2	0.30103	
4	0.60206	
Difference	0 30103	100% increase

Numbers	Logarithms	
5	0.69897	
10	1.00000	100% increase
Difference	0.30103	

Thus if the **logarithms** of the values rather than the original figures are plotted constant differences (rises or falls) will then equal constant percentage changes.

Since, however, a great deal of time and effort is required to convert the original data into the form of logarithms, a more convenient procedure is to arrange the scale so that the logarithms may be plotted directly by reference to a special scale.



The logarithms in the longer procedure may be plotted on the arithmetic scale in the usual fashion. Thus, if it is desired to plot the value 2 its logarithm is determined (0.30103) and this

value plotted. If, however, a scale is prepared in advance with the position 0.30103 marked 2, then the data may be plotted without previously determining the logarithms of the values.

The relation between a simple arithmetic scale and a scale corresponding to it but prepared for plotting logarithms is shown below:

If a logarithmic ruling is used on both the X and Y axis the paper is known as logarithmic, if used only on one axis it is semi-logarithmic.

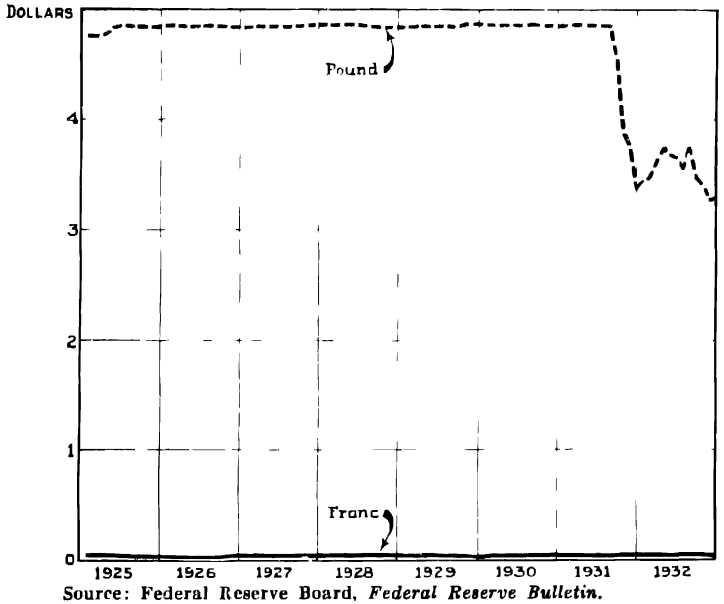
Since time is generally placed on the X axis an arithmetic ruling is used on this axis in semi-logarithmic paper while the logarithmic ruling is retained on the Y axis.

Characteristics of Logarithmic Charts

1. There is no zero or base line.

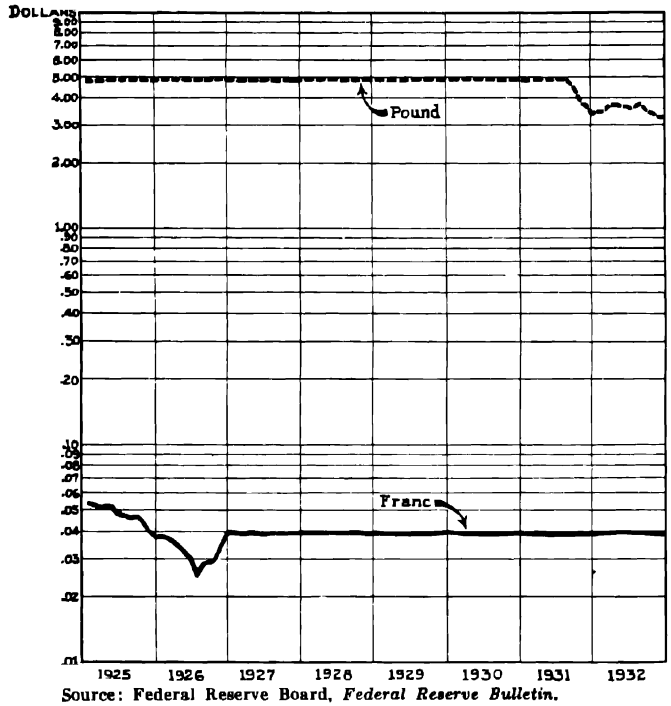
2. Semi-logarithmic charts have an arithmetic scale on the horizontal axis. Logarithmic charts are ruled logarithmically on both scales.

3. When plotted on logarithmic paper a geometric progression forms a straight line, since the logarithms of a geometric progression form an arithmetic progression.



Source: Federal Reserve Board, *Federal Reserve Bulletin.*

Fig. 31a—Exchange Rates on the Franc and the Pound Sterling, 1925—1932. (Plotted on Arithmetic Paper.)



Source: Federal Reserve Board, *Federal Reserve Bulletin.*

Fig. 31b—Exchange Rates on the Franc and the Pound Sterling, 1925—1932. (Plotted on semi-logarithmic paper.)

4. Equal rises or falls indicate equal percentage changes.

5. Equal slopes on a logarithmic chart denote equal rates of change.

Logarithmic Charts are Used:

1. To compare proportional rates of change.

2. To show the relationship between the two or more series which differ widely in amount.

   a. The unsatisfactory nature of arithmetic paper as compared with semi-logarithmic paper for this purpose can be seen from figure 31.

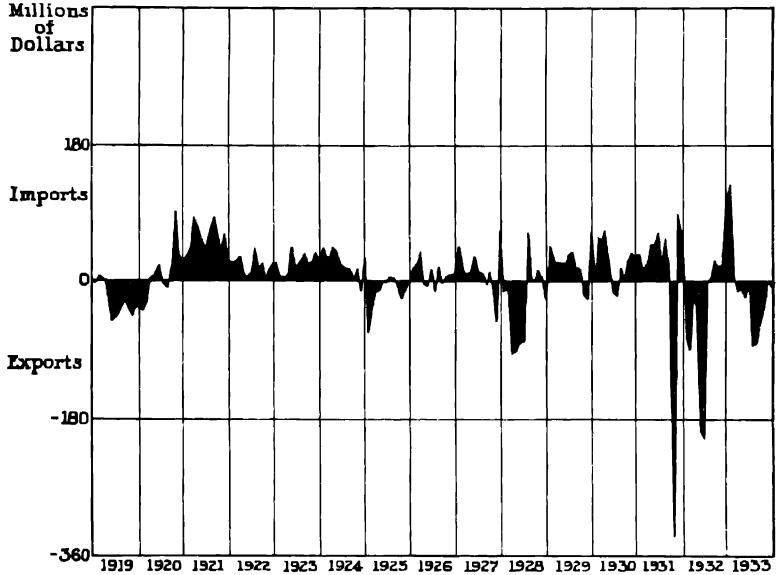
Special Types of Line Graphs

1. **Silhouette charts** are line graphs showing the positive and negative deviations from a zero or base line with the area between the zero or base line and the curve filled in (see figure 32).

Silhouette charts are constructed by plotting points indicating the actual deviations from the base line. The points are then connected and the area between the curve and the base line filled in.

2. **Band charts** are a form of line graph which show variations in the component parts as well as the total.

The chart is prepared by first plotting the variation in the largest component part of the total. This segment may then be shaded in or cross hatched. The next component part is then added to this first segment and the result plotted. This cumu-



Source: Standard Trade and Securities Service, *Statistical Base Book.*

Fig. 32—Gold Movements from the United States, 1919—1933. (Silhouette Chart).

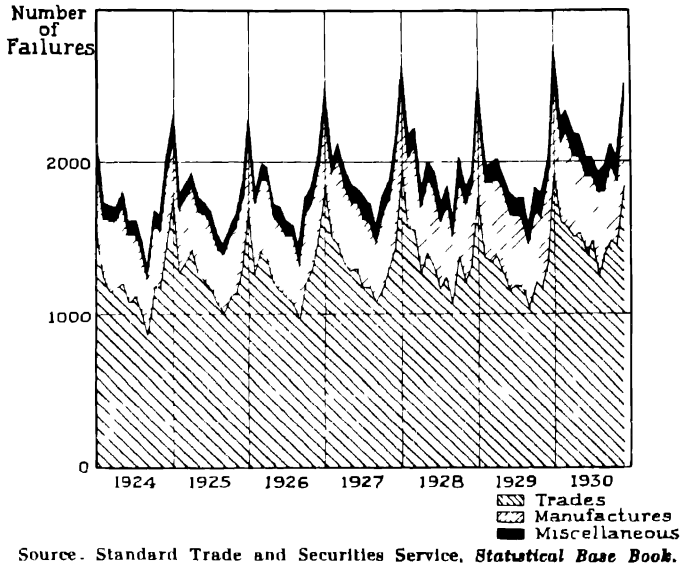
lative process is then continued until all of the component parts have been included. The variations in the top line will then represent variations in the total while the variations in the *width* of any segment will indicate the variations in that particular component part. Figure 33 illustrates this type of graph.

3. **High-low** graphs are a form of line graph which present not only the changes occurring over a period of time but the fluctuations occurring within each period (as day, week, month, etc.) as well, indicating the high and low values.

The high-low chart is constructed by first plotting the lowest value for a period and then the highest value for the same period. This procedure is continued until the end of the time covered by the

graph. The low point and the high point for each period are connected by means of heavy lines. These lines since they are closely spaced tend to take the appearance of an irregular band.

4. **The histogram,** also known as a rectangular frequency polygon is constructed from a frequency distribution in the following manner. Rectangles are erected using as the width the size of the class interval and as the height the frequency in each class interval. See page 3.



Source. Standard Trade and Securities Service, *Statistical Base Book.*

Fig. 33--Commercial Failures in the United States, by Types, 1924—1930. Band Chart.

Bar Charts

Bar charts visually contrast quantities by a comparison of bars of varying length but uniform width.

Bar charts may be subdivided into four types, namely:

1. Absolute
   a. Simple
   b. Subdivided
2. Percent
   a. Simple
   b. Subdivided

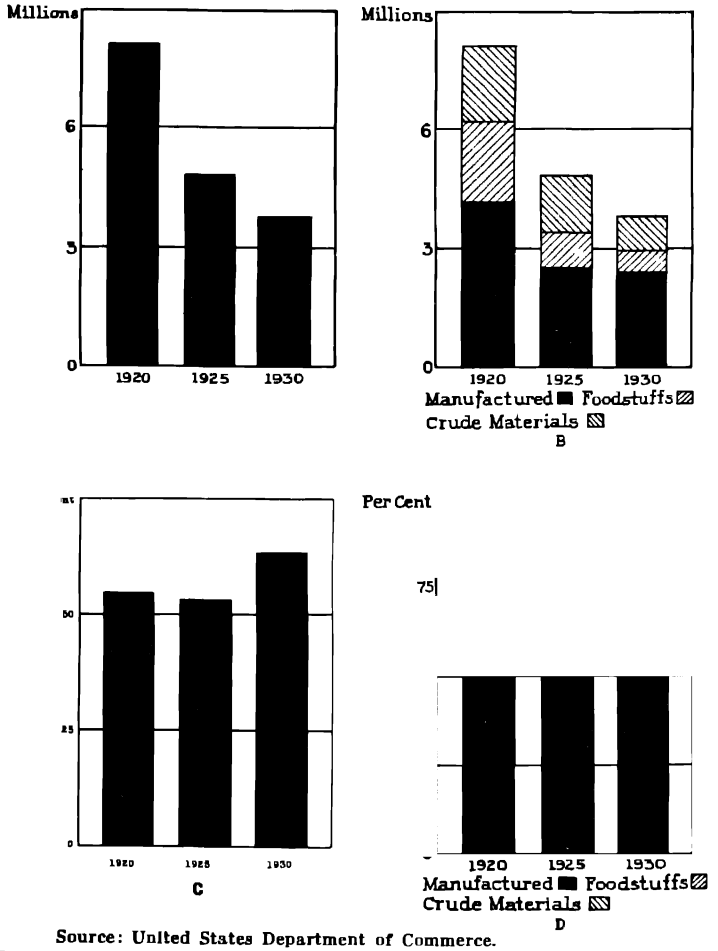
Simple Absolute Bar Charts

Rectangular bars of the same width are erected from the same base line to proportionate lengths based on absolute or actual

data. The bars may be set up either horizontally or vertically; however, when the scale involves time the vertical type of bar is indicated. See graph A, figure 34.

Subdivided Absolute Bar Charts

The bars are subdivided according to the size of each component. The components of each bar are arranged in similar order with the largest subdivisions at the base. The figures may vary to such an extent that the largest subdivision may not re-



Source: United States Department of Commerce.

Fig. 34—Exports from the United States, 1920, 1925, and 1930. Shown as various forms of bar charts. (Chart C is for Manufactured exports as percent of total.)

tain its position, nevertheless the order of arrangement should remain fixed. Subdivided charts are cumulative in that each subdivision in plotting is added to the total of the subdivisions below it (see graph B, figure 34).

Simple Percentage Bar Charts

The bars are constructed in a similar fashion to the method used in simple absolute bar charts except that the lengths of the bars represent percentage values (see graph C, figure 34).

Subdivided Percentage Bar Charts

Rectangular bars of the same width and the same length are constructed on the same base. The length represents 100 per cent. Each bar is divided into segments, the size of each segment being dependent on the percent of the total figure which each subdivision represents.[1]

The subdivisions of each bar are arranged in the same order of presentation with the largest percentage at the base (see graph D, figure 34).

A special type of subdivided percentage bar chart is that which makes use of a single bar. The single bar is used when interest is centered on the component parts of a single total. The entire length of the bar represents 100 percent and each segment is represented in order of size from left to right.

Pictorial Bar Charts

Bar charts may be constructed in pictorial form. Pictures of different heights may be used for comparative purposes. Thus to represent the gold holdings of the United States Treasury at different periods stacks of coins of varying heights corresponding to the values they represented may be used.

Loss and Gain Bar Charts

This type of bar chart is constructed by having the bars extended from a zero line. If the bar chart is constructed horizontally the bars representing losses extend to the left, profits to the right. If the bar chart is constructed vertically the bars above the horizontal normal line represent profits, the bars below the normal line represent losses.

Area Diagrams

The area diagram contrasts quantities by comparing figures with varying areas. Area diagrams may be of many types the simplest making use of geometric figures (such as circles and squares). Area diagrams are of two types. In the first type total areas of different sizes may be contrasted by varying the sizes of the figures. In the second type subdivisions of a single area may be compared.

[1] In subdivided bar charts the number of subdivisions should be as few as possible.

The most usual type of area diagram is the **pie chart** (see figure 35).

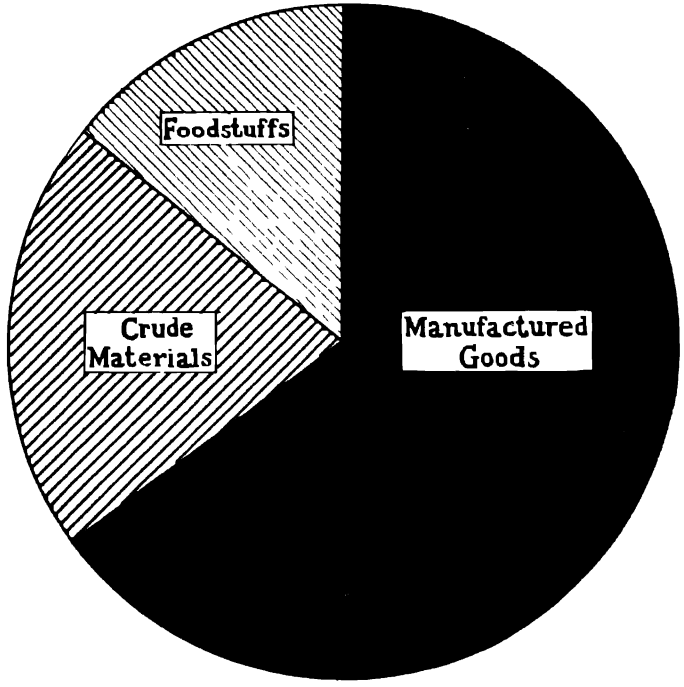
Pie Chart

Definition

A pie diagram is a chart of circular shape broken into subdivisions. The size of the section indicates the proportion of each component part to the whole.

Construction

1. Let the circle equal 100 percent
2. Each circle is divided into 360°.
3. $\therefore$ each percent $= \frac{360°}{100}$ or 3.6°



Source: United States Department of Commerce

Fig. 35—Exports from the United States, by Economic Classes, 1930.

Characteristics

1. The arrangement of the size of sectors is generally clockwise according to size.

2. A uniform arrangement of sectors must be made in comparing charts.

3. Wherever possible wording and percentages should be placed horizontally on the sector.

4. If shading, cross hatching, colors, etc. are used in place of wording on sectors, a legend should be constructed to indicate their meaning.

5. The effectiveness of the pie chart is enhanced by cross hatching, colors, shades, etc.

6. A pie chart should have a minimum of sectors.

7. The pie chart is difficult to construct accurately.

8. It is difficult to estimate visually with any degree of accuracy the proportionate size of the sectors of a pie chart where percentages are not indicated.

Solid Diagrams

Solid diagrams consist of geometric forms (cubes, spheres, cylinders, etc.) or irregular figures constructed to illustrate comparisons of magnitudes through comparison of volumes at the figures (see figure 36). The volumes of the figures in a solid diagram as compared and not the heights or lengths of the figures. The solid diagram makes accurate comparisons difficult and for this reason it should not be used if some other method of illustration is possible.

Map Graph

Function: The map graph presents in pictorial form the facts in a geographic distribution.[1]

Construction: Map graphs are of five major types:

1. Shaded
2. Cross Hatched
3. Dotted
4. Colored
5. Pin
   a. Tacks
   b. Pins
   c. Flags

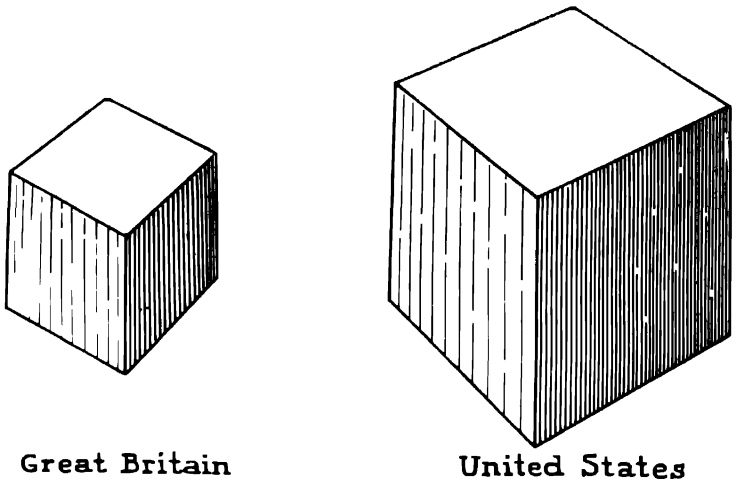
Shaded Maps

The proportionate quantities for particular areas may be indicated by using various degrees of shading ranging from solid black to white.

Cross Hatched Maps

Cross hatching may be used to indicate varying quantities by varying the proportions of black and white space.

[1] A geographic distribution is known also as a "spatial" distribution



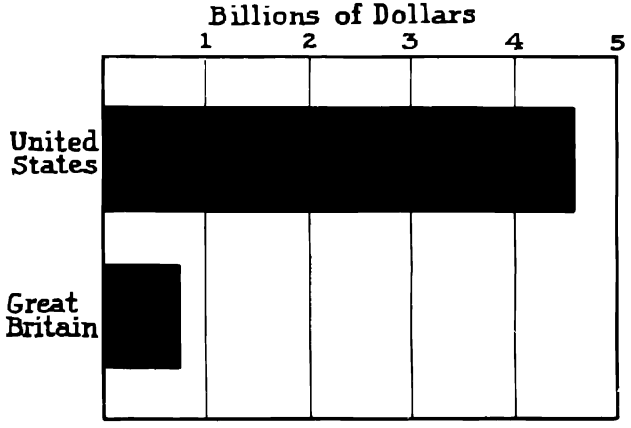


Fig. 36—Gold Stock of the United States and Great Britain, December 31, 1930. Shown in Solid Diagram and Bar Chart Form..

Dotted Maps

Dots (circular areas) on maps are used primarily in two ways:

a. **Dots of similar size** are placed on a map the primary purpose of which is to indicate the ***density*** of the numbers, etc., in an area by varying the number of these dots.

b. **Dots of proportional sizes** are placed on a map to indicate the total number or sizes in an area

c. **Dots of a fixed size** each with the same assigned value may be valued in number to indicate the various quantities for each area.

Care must be taken in denoting relative sizes since comparisons must be made by varying the area of the dot.

Colored Maps are constructed by using

a. Various colors to indicate variations in sizes, etc.

A variety of colors should not be used to indicate relative values since an individual color is of no greater value than any other color to the observer.

b. Various degrees of the same color to illustrate relative positions of different areas. The difficulty with using a single color scheme is that there is a limited number of shades which can be satisfactorily used.

Map-Tack System

Maps using tacks, flags, etc., are used for a large number of purposes in indicating relative sizes and also densities in a geographic area.

The heads of the pins or tacks may be of various colors, sizes, and shapes and thus extend their flexibility for the indication of sizes, locations, routes, etc.

ADDITIONAL BIBLIOGRAPHY*

ARKIN, HERBERT & COLTON, RAYMOND R., *Graphs.* Harper and Bros., New York, Second Edition, 1938.

BOWLEY, ARTHUR L., *Elements of Statistics*, pp. 125–72. P. S. King & Son, London, 1901.

ODELL, C. W., *Educational Statistics*, pp. 36–61. Century Co., New York, 1925.

OTIS, ARTHUR S., *Statistical Method in Educational Measurements*, pp. 30–35; 53–67. World Book Co., Yonkers, New York, & Chicago, Illinois, 1926.

SUTCLIFFE, WILLIAM G., *Statistics for the Business Man*, pp. 18–62. Harper & Bros., New York, 1930.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

CHAPTER XIX

SPECIAL TECHNIQUES IN EDUCATION, PSYCHOLOGY AND BIOLOGY

Special Techniques in Education and Psychology

Standard Scores[1]

In order to compare the results of two or more tests given to a number of students the tests should be of equal difficulty. If they are unequal in difficulty the average grades for each examination may differ widely, as will the resulting dispersion of the grades. The difficulty in the comparison of grades can be seen in the following distribution of the results for two examinations:

Student Number	Examination Number 1	Examination Number 2
1	100	100
2	99	85
3	99	70
4	98	68
5	97	65
6	96	60
7	90	60
8	90	60
9	87	50
10	80	45

The grades or scores may be standardized by converting each grade into a deviation from its respective arithmetic mean and dividing by the standard deviation to allow for differing average attainment and dispersion of grade.

$$z = \frac{x}{\sigma}$$

where

z = standard score

x = deviation of given score from mean.

σ = standard deviation of original grade.

The grades on both examinations will average to the same value—zero (since $\Sigma\ (x) = 0$)—and since each grade has been made relative to its standard deviation[2] the standard scores on both tests will have the same degree of dispersion.

[1] Compare Kelley, T. L., *Statistical Methods*

[2] If it is desired to have a fixed average score of 50 in each test the formula $s = 50 + \frac{z}{\sigma} 10$ may be used

If a pupil is consistent the standard scores attained on two tests measuring different traits, such as silent reading and arithmetic, should be equal:

$$z_1 = z_2$$

A large divergence in the value of these standard scores indicates a definite idiosyncrasy of the pupil considered.

The Coefficient of Reliability

The reliability of any test is measured by the similarity of results attained when the same test is given a number of times. If the reliability of the test or other measuring instrument is perfect, exactly similar results (allowing for chance variation) should be obtained when the test is given twice or the coefficient of correlation between the two sets of scores will be 100.

The coefficient of correlation of the scores secured from two applications of the same measuring instrument is thus a coefficient of reliability.

However it is usually not practical to give the same examination or the same form of examination twice. In place of this it is preferable to use the coefficient of correlation between the scores attained by dividing the questions into two parts. For this purpose the odd numbered questions (1, 3, 5, 7 etc.,) may be used as one set and the even numbered questions (2, 4, 6, 8 etc.,) as the other. For this purpose the distribution of questions between the two sets must be at random or at least the two sets of questions must be of equal difficulty.

The increase in the reliability of a test secured by lengthening it or repeating it a number of times in the same form may be computed from the Spearman-Brown formula

$$r_n = \frac{n\, r_{11}}{1 + (n - 1)\, r_{11}}$$

where

r_n is the increased reliability coefficient resulting from either increasing the length of the test n times or repeating it n times.

r_{11} is the coefficient of reliability for the original test.

The Intelligence Quotient

The scores on various intelligence tests given an individual will increase with his age. It is necessary, therefore, to relate the *mental age* as indicated by the intelligence tests to his *chronological age.*

$$I.Q. = \frac{MA}{CA}$$

where

$I.Q.$ = Intelligence Quotient

MA = Mental age

CA = Chronological age.

For statistical purposes the *normal* intelligence quotient for any age will then be 100.

In order to calculate the value of the intelligence quotient the chronological age is increased until it reaches a maximum at which it is retained. Otis[1] argues that the maximum age should be 18 rather than the generally used maximum of 16 years.

Subject Quotients and Ratios

Pupil accomplishment in a particular subject varies according to chronological age. To obtain a true picture of relative ability in a given subject it is necessary to compare "subject age" to chronological age.

$$\text{Arithmetic Quotients} = \frac{\text{Arithmetic Age}}{\text{Chronological Age}}$$

$$\text{Reading Quotients} \quad \frac{\text{Reading Age}}{\text{Chronological Age}}$$

$$\text{Any subject Quotients} \quad \frac{\text{Any subject Age}}{\text{Chronological Age}}$$

The average of a pupil's subject quotients is known as his **educational quotient.**

If the mental age is used in the quotients outlined above the result is known as a **subject ratio,** etc.

$$\text{Subject ratio} \quad \frac{\text{Subject Age}}{\text{Mental Age}}$$

The **accomplishment ratio** may then be obtained by averaging the subject ratios.

Special Techniques in Biology

Index of Abmodality

A deviation from average or type is of little significance unless the deviation is related to the customary dispersion of the data. A deviation of two inches from the average height of a man of a certain age is of little import unless compared to the ordinary or usual dispersion of man heights of the same age. The same principle would hold true in the deviation of an inch from the average length of a squirrel of a specified age, etc., etc.

In order to take the dispersion of the data into consideration the deviation from the mean may be related to the standard deviation of the data.[2]

$$\frac{x}{\sigma}$$

This measure is known to biologists as the **index of abmodality.**

[1] Otis, Arthur S., *Statistical Method in Educational Measurement*, 1926, p 150.

[2] In the field of education this is known as the standard score (see page 111)

The index of abmodality for an essentially normal distribution indicates the number of standard deviations the given value is from the mean. Thus if the index attains a certain value it may be further interpreted in light of the previous discussion on the normal curve (see chapter XI).

It is known that a deviation larger than 3 standard deviations from the mean (in any one direction plus or minus) will occur less than 2 times in 1000. This knowledge is obtained from the demonstrable fact that the area within 3 standard deviations from the mean will include 49.87% of the cases (on one side only of the mean), and therefore only .15% of the cases will be larger than the value of the index of abmodality if it equals 3. Any given value of the index of abmodality may thus be interpreted with the aid of the normal curve area table.

Coefficient of Heredity

When the coefficient of correlation is applied to the measurement of the association between a specific characteristic of a parent and the same characteristic of an offspring it is known as the **coefficient of heredity.**

The coefficient of heredity between fathers and offspring is assigned the symbol r_1 between mothers and offspring r_2.

Coefficient of Assortative Mating

When the coefficient of correlation is used to measure the association between a specified characteristic of fathers and the same characteristic of mothers it is known as the coefficient of **assortative mating.** The symbol assigned to this coefficient is r_3.

Variability of Offspring

The variability (standard deviation) of a group of offspring from particular parents may be determined from the following formula:

$$\sigma_{3\,12} \quad \sigma_3 \sqrt{1 - \frac{2r^2_1}{1 + r_3}}$$

where

$\sigma_{3\,12}$ = standard deviation of an array of offspring.

σ_3 = standard deviation of offspring in general.

r_1 = coefficient of heredity between offspring and parents, assuming parents to be equipotent ($r_1 = r_2$).

r_3 = Coefficient of assortative mating.

Abmodality of Offspring

The average abmodality of a group of offspring from parents of fixed characteristics may be computed from the formula:

$$h_3 = \frac{r_1\,\sigma_3}{(1 + r_3)\,\sigma_1}\left(h_1 + \frac{\sigma_1}{\sigma_3}\,h_2\right)$$

where

h_3 = deviation of mean of given characteristic of offspring from mean of characteristic of all offspring.

h_1 = deviation (abmodality) of father.

h_2 = deviation of mother.

σ_1 = standard deviation of characteristic in fathers in general.

σ_2 = standard deviation of characteristic of mothers in general.

σ_3 = Standard deviation of offspring in general.

r_1 = coefficient of heredity (assuming $r_1 = r_2$).

r_2 = coefficient of assortative mating.

Vital Statistics

Data relative to deaths, births, and sickness are significant only if considered in relation to the size and kind of population from which they were drawn. Thus the fact that 2,000 deaths were recorded in a year in a particular city is of no significance unless the population of the city is known. If the 2,000 deaths were recorded in New York City with a population of approximately 7,400,000 an entirely different significance would be attached to such a record than if it were recorded in a city of 25,000 population

A **rate** is an expression of the number of times a specific kind of event occurs in a given population in relation to the total number in the population exposed to the possibility of its occurrence. This may be expressed in the form of a formula as:

$$\text{Rate} \quad \frac{a}{a+b}$$

where

a = number of times event appears in the population

b = number of times event does not appear in the population

The resulting value is in decimal form but is generally multiplied by 100, 1,000, 100,000 or 1,000,000 to give the result as percent (per 100), per 1000, per 100,000 or per million.

A **ratio** expresses the relation of occurrence of a given kind of event to the occurrence of other events or of one kind of data to another. In formula form this is:

$$\text{Ratio} = \frac{a}{c}$$

where

a = number of times event occurs

c = number of times another event occurs

Vital statistics make use of birth, death and morbidity rates. These rates are important also in medical and actuarial statistics. Birth, death and morbidity rates may be classified as follows:[1]

[1]This classification is after Pearl, Raymond in *Medical Biometry and Statistics*.

A—Mortality Rates (Death Rates)
1. Observed
   a. Crude death rates
   b. Specific death rates
2. Theoretic Death Rates
   a. Standardized death rates
   b. Corrected death rates

B—Natality Rates (Birth Rates)
1. Observed
   a. Crude Birth Rates
   b. Specific Birth Rates
2. Theoretic Birth Rates
   a. Standardized Birth Rates
   b. Corrected Birth Rates

C—Morbidity Rates
1. Observed
   a. Crude
   b. Specific

Crude Death, Birth and Morbidity Rates

The crude death, birth or morbidity rates are merely the total number of deaths, births, or cases of sickness divided by the total population

$$\text{Crude Death Rate} = \frac{D}{P}$$

$$\text{Crude Birth Rate} = \frac{B}{P}$$

$$\text{Crude Morbidity Rate} = \frac{M}{P}$$

where

D = number of deaths
B = number of births
M = number of persons sick

Specific Death Rate

Although these rates must be specified as to time and place they are crude in that they do not include specifications as to age or sex. When such specifications are made the rate is known as the specific death, birth, or morbidity rate.

Specific Death, Birth, or Morbidity Rate =

$$\frac{D' \text{ or } B' \text{ or } M'}{P'}$$

where

D' = deaths in a specified class of population

B' = births in a specified class of population

M' = number of persons sick in a specified class of population

P' = total number of persons in the specified population group.

The specific death rates at various ages will of course vary greatly.

If it is assumed that 100,000 persons are born at the same instant a hypothetical table showing the number of survivors, the number dying, the rate of mortality and the stationary population at each age interval can be constructed. A table of this kind is known as a life table.[1]

From the life table a stationary life table may be prepared showing the number of persons per million of each yearly interval of age. Table 42 is such a table.

Standardized Death Rates

Due to various factors such as immigration, type of community, etc. the actual age distribution in one location may differ greatly from that in another community making it impossible to directly compare the crude death rates for all ages for the two localities. An allowance must be made for the difference. This is accomplished by means of standardized and corrected death rates.

The standardized death rate is obtained by applying the specific death rate obtained from the general population or a life table to the actual age distribution of the given population. The rate obtained is the rate that would exist if the hypothetical specific rate existed with the actual distribution of age.

$$\text{Standardized Death Rate} = \frac{\Sigma\,(p\,q)}{\Sigma\,(q)}$$

where

p actual population for each age

q specific death rate from life table

A comparison of this rate to the death rate of the standard population (from the life table) gives use to a correction factor.

$$\text{Correction Factor} = \frac{R}{R'}$$

where

R death rate from life table

R' standardized death rate

Multiplying the crude death rate by this correction factor will make an allowance for the different age distributions.

[1] See Glover, J W , United States Life Tables 1890, 1901, 1910 and 1901-1910 Bureau of Census, Washington, 1921

Table 42—Stationary Life Table Population of 1,000,000 Persons. Number Living in Each Yearly Interval of Age.

Age Interval	Persons Per Million in Current Age Interval	Age Interval	Persons Per Million in Current Age Interval	Age Interval	Persons Per Million in Current Age Interval
0– 1	17,841	35–36	14,146	70– 71	6,373
1– 2	16,916	36–37	14,031	71– 72	5 979
2– 3	16,612	37–38	13,912	72– 73	5,597
3– 4	16,448	38–39	13,791	73– 74	5,178
4– 5	16,338	39–40	13,667	74– 75	4,776
5– 6	16,255	40–41	13,540	75– 76	4,375
6– 7	16,186	41–42	13,411	76– 77	3,978
7– 8	16,127	42–43	13,278	77– 78	3,589
8– 9	16,078	43–44	13,141	78– 79	3,210
9–10	16,036	44–45	13,000	79– 80	2,843
10–11	15,998	45–46	12,854	80– 81	2,490
11–12	15,962	46–47	12,702	81– 82	2,152
12–13	15,927	47–48	12,545	82– 83	1,835
13–14	15,890	48–49	12,383	83– 84	1,546
14–15	15,851	49–50	12,216	84– 85	1,287
15–16	15,808	50–51	12,045	85– 86	1,058
16–17	15,761	51–52	11,867	86– 87	859
17–18	15,708	52–53	11,683	87– 88	687
18–19	15,650	53–54	11,489	88– 89	541
19–20	15,586	54–55	11,284	89– 90	418
20–21	15,516	55–56	11,067	90– 91	318
21–22	15,441	56–57	10,836	91– 92	236
22–23	15,363	57–58	10,592	92– 93	172
23–24	15,282	58–59	10,336	93– 94	123
24–25	15,200	59–60	10,069	94– 95	86
25–26	15,117	60–61	9,791	95– 96	59
26–27	15,032	61–62	9,501	96– 97	39
27–28	14,946	62–63	9,199	97– 98	26
28–29	14,857	63–64	8,884	98– 99	17
29–30	14,765	64–65	8,556	99–100	10
30–31	14,671	65–66	8,217	100–101	6
31–32	14,573	66–67	7,868	101–102	4
32–33	14,472	67–68	7,508	102–103	2
33–34	14,367	68–69	7,139	103–104	1
34–35	14,259	69–70	6,760	104–105	1

Source: Pearl, Raymond, *Medical Biometry and Statistics*, page 259.

Corrected Death Rates

The corrected death rate is obtained by using the specific death rates of the locality and hypothetical age distribution of the life table. This computation places the rates on a strictly comparable basis in-so-far as the age distribution is concerned.

$$\text{Corrected Death Rate} \quad \frac{\Sigma\,(p'\,q')}{\Sigma\,(p')}$$

where

p' = population in age group from life table

q' = actual specific death rates

Corrections have thus been made for varying age distributions. In a similar fashion corrections may be made for other differences such as sex, race, etc. Rates other than death rates may be corrected in the same manner.

Production

Quality Control

The quality of a given product may be defined as its conformity to given standards or specifications. A manufactured product exhibits a certain amount of variation in its conformity to specifications, no matter how carefully guarded the process may be, due to innumerable chance causes. As long as only these chance, uncontrollable factors are the cause of variation the quality is said to be controlled. However, as soon as some controllable cause enters to cause variation, control is no longer had and the product departs from conformity to standard.

The problem of quality control is thus essentially the determination of the entrance of factors other than those which from an economic viewpoint should be left to chance. Dr. Shewhart[1] suggests the following criteria to determine whether the variations exhibited by any measured characteristics of a manufactured product are due to chance or some controllable factor or in other words the existence of or lack of "control." [2]

Criterion I

1. Divide the data into m rational (groups such as week, day, plant, etc.) of n items each.
2. Determine a statistic such as $\overline{X}$ or σ for the entire group.
3. Represent this value as a horizontal line on a graph.
4. Compute the standard error of the statistic,

$$\sigma_p = \sqrt{N\,p\,q}$$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}}$$

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} \text{ etc.}$$

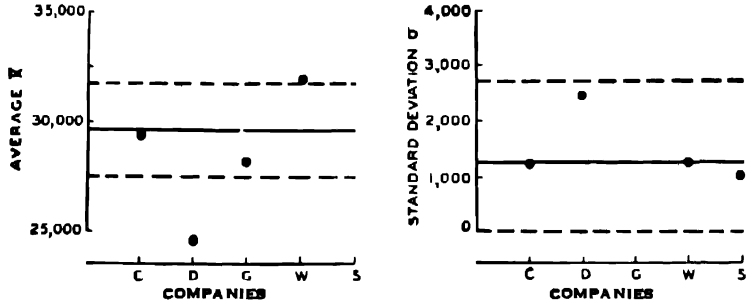
When n is small the theory of small samples (see pages 129-130) must be used.

5. Measure off a zone of 3 standard errors on either side of the horizontal line. The resulting diagram is a control chart.

[1] W. A. Shewhart, *Economic Control of Quality of Manufactured Product*, D. Van Nostrand, New York, 1931.

[2] The discussion of this topic is of necessity very brief. For a detailed and complete discussion which is essential to a thorough understanding the reader is referred to Dr. Shewhart's book

6. The statistics for each group (X, p, σ etc.) may now be located on the control chart. If any of the points fall outside of the control zone there is an indication of the presence of assignable causes of variability which should be investigated.



From Shewhart, W. A. *Economic Control of Quality of Manufactured Product*, p. 313.

Fig. 37. Control Chart for Small Samples Showing Lack of Control

Criterion II

1. Divide the data into m rational groups of n items each.
2. Compute d

$$d = \frac{n}{n-1}\bar{\sigma}^2 - \frac{m}{m-1}n\sigma_x^2$$

where

$\bar{\sigma}^2$ = average σ^2 for all groups

$\sigma_{\bar{x}}^2$ = variance (square of standard deviation) of the means of the groups.

For a Bernoulli distribution—a constant system of causes —d will equal zero. However due to sampling fluctuations a value as great as $3\sigma_d$ may arise with the probability that 99.7 chances out of 100 that no greater value will occur where

$$\sigma_d = \left[\sqrt{\frac{2(mn-1)}{m(m-1)(n-1)}}\left(\frac{n}{n-1}\bar{\sigma}^2\right)\right]$$

If a value of d greater than $3d$ is secured it is an indication that the samples are not drawn in a Bernoulli fashion or that the system of causes is not controlled. Indication is had that control is lacking.

Criterion III

1. Determine a suspected factor which might be the only assignable cause of the variability and two sets of variables in which may be caused by the suspected factor. One of the variables should be controlled.

2. Secure the coefficient of correlation (r) between these two variables.
3. If r is greater than $3\sigma_r$ indicating a significant correlation, the suspected factor may be said to be the assignable cause.

Criterion IV

1. Obtain n observations and calculate some statistic (such as $\overline{X}$ or σ).
2. Choose some factor which may or may not be an assignable cause of variation and select n additional observations under conditions where it is known that they cannot be affected by this factor.
3. By means of the standard error of the difference between the two computed statistics determine whether the difference is significant. If the difference between the statistics is greater than three times the standard error of the difference, the suspected cause is the assignable cause as when:

$$\overline{X}_1 - \overline{X}_2 \qquad 3\sqrt{\frac{1}{n}(\sigma_1^2 + \sigma_2^2)}$$

$$\sigma_1 - \sigma_2 \qquad 3\sqrt{\frac{1}{2n}(\sigma_1^2 + \sigma_2^2)}$$

Criterion V

1. Fit an appropriate frequency curve to the grouped data by calculating $\overline{X}$, σ and the skewness k and using

$$\frac{1}{\sigma\sqrt{2\pi}}\left[1 - \frac{k}{2}\left(x - \frac{1}{3}\frac{x^2}{\sigma_3}\right)\right] \quad \frac{-x^2}{2\sigma^2}$$

2. Test the theoretical distribution for goodness of fit through the χ^2 test. If the fit is poor (P is less than .001) this is an indication of lack of control.

ADDITIONAL BIBLIOGRAPHY*

Techniques in Education and Psychology

FOSTER, S., *Experiments in Psychology*, Henry Holt & Co., New York, 1923.

HINES, H. C., *A Guide to Educational Measurements*. Houghton Mifflin Co., Boston, 1923.

KELLEY, T. L., *Interpretation of Educational Measurements*. World Book Co., Yonkers, New York, 1927.

MCCALL, W. A., *How to Experiment In Education*. Macmillan Co., New York, 1923.

* For readings in standard Statistics textbooks, see the QUICK REFERENCE TABLE TO STANDARD TEXTBOOKS following Table of Contents.

OTIS, A. S., *Statistical Method in Educational Measurements.* World Book Co., Yonkers, New York, 1926.

Techniques in Biology

DAVENPORT, C. B., *Statistical Method with Special Reference to Biological Variations.* 3rd Revised Edition. John Wiley & Sons, New York, 1904.

DAVENPORT, E., *Principles of Breeding.* Ginn & Co., Boston, 1907.

Techniques in Vital Statistics

GLOVER, J. W., *United States Life Tables, 1890, 1901, 1910, and 1901-1910,* Bureau of Census, Washington, 1921.

PEARL, RAYMOND, *Medical Biometry and Statistics.* W. B. Saunders, Philadelphia, 1930.

Techniques in Production Control

SHEWHART, W. A., *Economic Control of Quality of Manufactured Product,* D. Van Nostrand Co., New York, 1931.

APPENDIX

LIST OF FORMULAS

ANALYSIS OF THE FREQUENCY DISTRIBUTION

The following reference list includes all of the formulas contained in this volume and many others dealt with in more extended texts. The number appearing on the right side of the page are the page references to this volume. The names appearing in that column are the names of the authors discussing the formulas not covered in this book. For the full title of the volumes see the appended list of references.

Arithmetic Mean

From Ungrouped Data

$$\overline{X} = \frac{\Sigma (X)}{N} \qquad 11$$

From Grouped Data

A. Long Method

$$\overline{X} = \frac{\Sigma (f \times M.P.)}{N} \qquad 13$$

B. Short (unit deviation) Method

$$\overline{X} = \overline{Z} + \frac{\Sigma (fd)}{N} \qquad 15$$

C. Short (group deviation) Method

$$\overline{X} = \overline{Z} + \frac{\Sigma (fd')}{N} C \qquad 16$$

Median

From Grouped Data

$$\text{Median} = L + \frac{i}{f} C \qquad 20$$

Mode

From Grouped Data

A. Moments of Force Method

$$\text{Mode} = L_{mo} + \frac{f_a}{f_a + f_b} C \qquad 23$$

B. Empirical Method

$$\text{Mode} = \text{Mean} - 3\,(\text{Mean} - \text{Median}) \qquad 24$$

C. Another Method

$$\text{Mode} = \overline{X} - (\chi)(\sigma) \qquad 152$$

Geometric Mean

From Ungrouped Data

$$G_m = \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdot \ldots \ldots X_n} \qquad 26$$

$$\log G_m = \frac{\log X_1 + \log X_2 + \log X_3 + \ldots\ldots\ldots + \log X_n}{N} \qquad 26$$

From Grouped Data

$$\log G_m = \frac{\Sigma\,(f \log \text{M.P.})}{N} \qquad 27$$

Quadratic Mean

$$Q_m = \sqrt{\frac{\Sigma\,(X^2)}{N}} \qquad 27$$

Harmonic Mean

$$\frac{1}{H_m} = \frac{\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} \quad \ldots\ldots\ldots\ldots + \frac{1}{X_n}}{N} \qquad 27$$

Measures of Dispersion

Mean Deviation

Ungrouped Data

$$MD = \frac{\Sigma\,|x|}{N} \text{ or } \frac{\Sigma\,|d|}{N} \qquad 32$$

From Grouped Data

$$MD' = \frac{\Sigma\,(fd')}{N} + \frac{(N_S - N_L)c}{N} \qquad MD = MD' \times C \qquad 33$$

$$M.\,D. + \frac{\Sigma\,(fd) + (N_S - N_L)\,c}{N}\,C \qquad 33$$

$$MD' = \frac{\Sigma\,(fd)}{N} + \frac{(N_a + N_b)c + f_m\,(.25 + c^2)}{N} \qquad \text{Rietz}$$

Standard Deviation

From Ungrouped Data

$$\sigma = \sqrt{\frac{\Sigma (x^2)}{N}} \qquad 34$$

$$\sigma = \sqrt{\frac{\Sigma (X^2)}{N} - \left(\frac{\Sigma X}{N}\right)^2} \qquad 34$$

From Grouped Data

Long Method

$$\sigma = \sqrt{\frac{\Sigma f(x^2)}{N}} \qquad 36$$

Short (unit deviation) Method

$$\sigma = \sqrt{\frac{\Sigma f (d^2)}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \qquad 36$$

Short (group deviation) Method

$$\sigma = C \sqrt{\frac{\Sigma f (d')^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \qquad 36$$

Correction of Standard Deviation for Grouping

$$\sigma'^2 = (\sigma^2 - 1/12)\ C^2 \qquad 37$$

Charlier Check for Computation of Standard Deviation

$$\Sigma f (d' + 1)^2 = \Sigma fd'^2 + 2 \Sigma (fd') + N \qquad 38$$

Quartile Deviation (Semi-Interquartile Range)

$$QD = \frac{Q_3 - Q_1}{2} \qquad 40$$

Coefficient of Variation

$$V = \frac{\sigma}{\bar{X}} 100 \qquad 41$$

$$V_{AD} = \frac{AD}{\text{Median (or mean)}} \qquad 41$$

$$V_Q = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \qquad 41$$

Measures of Skewness

Coefficient of Skewness

$$S_K = \frac{\text{Mean} - \text{Mode}}{\sigma} \qquad 42$$

$$S_K = \frac{3\ (\text{Mean} - \text{Median})}{\sigma} \qquad 42$$

$$S_K = \frac{(Q_3 - \text{Median}) - (\text{Median} - Q_1)}{QD} \qquad 42$$

Other Measures of Skewness

$$\alpha_3 = \frac{\mu_3}{\sigma_3} = \sqrt{\beta_1} \qquad 151$$

$$\chi = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2\,(5\,\beta_2 - 6\,\beta_1 - 9)} \qquad 152$$

Further Analysis of Frequency Distribution

Kurtosis

$$\beta_2 - 3 \qquad 151$$

$$\frac{\beta_2 - 3}{2} \qquad 151$$

Moments

About Arbitrary Origin

I. The first moment

$$\nu_1 = \frac{\Sigma(fd)}{} \qquad 148$$

II. The second moment

$$\nu_2 = \frac{\Sigma f(d^2)}{N} \qquad 148$$

III. The third moment

$$\nu_3 = \frac{\Sigma f(d^3)}{N} \qquad 148$$

IV. The fourth moment

$$\nu_4 = \frac{\Sigma f(d^4)}{N} \qquad 148$$

About Arithmetic Mean

$$\mu_1 = \frac{\Sigma f(x)}{N} = 0 \qquad 148$$

$$\mu_2 = \frac{\Sigma f(x^2)}{N} \qquad 148$$

$$\mu_3 = \frac{\Sigma f(x^3)}{N} \qquad 148$$

$$\mu_4 = \frac{\Sigma f(x^4)}{N} \qquad 148$$

$$\mu_2 = \nu_2 - \nu_1^2 \qquad 149$$

$$\mu_3 = \nu_3 - 3\ \nu_1\ \nu_2 + 2\ \nu_1^3 \qquad 149$$

$$\mu_4 = \nu_4 - 4\ \nu_1\ \nu_3 + 6\ \nu_1^2 \nu_2 - 3\,\nu_1^4 \qquad 149$$

Sheppard's Corrections for Grouping

I. Corrected First Moment (in class intervals)

$$\mu_1' = 0 \qquad 149$$

II. Corrected Second Moment (in class intervals)

$$\mu_2' = \mu_2 - 1/12 \qquad 149$$

III. Corrected Third Moment (in class intervals)

$$\mu_3' = \mu_3 \qquad 149$$

IV. Corrected Fourth Moment (in class intervals)

$$\mu_4' = \mu_4 - \tfrac{1}{2}\ \mu_2 + 7/240 \qquad 149$$

μ_2' (in original units) $= C^2\ \mu_2'$ (in class interval units) 149

μ_3' (in original units) $= C^3\ \mu_3'$ (in class interval units) 149

μ_4' (in original units) $= C^4\ \mu_4'$ (in class interval units) 149

Curve Criterion (Measure of Skewness)

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad 151$$

Curve Criterion (Measure of Kurtosis)

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \qquad 151$$

Curve Criterion

$$\kappa = \frac{\beta_1 (\beta_2 + 3)^2}{4\ (4\beta_2 - 3\beta_1)\ (2\beta_2 - 3\beta_1 - 6)} \qquad 151$$

Time Series Analysis

Straight Line

$$Y = a + bX \qquad 54$$

"Normal" Equations for Straight Line

$$\begin{aligned} \text{(I)}\quad & \Sigma(Y) = Na + b\Sigma(X) \\ \text{(II)}\quad & \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2) \end{aligned} \qquad 55$$

Simplified "Normal Equations for Straight Line
(Origin at midpoint of data)

$$\begin{aligned} \text{I}\quad & \Sigma(Y) = Na \\ \text{II}\quad & \Sigma(XY) = b\Sigma(X^2) \end{aligned} \qquad 58$$

Potential Equations

$Y = a + bX + cX^2$ (Second degree parabola)
$Y = a + bX + cX^2 + dX^3$ (Third degree parabola) 64, 93
$Y = a + bX + cX^2 + dX^3 + eX^4$ (Fourth degree parabola)
$Y = a + bX + cX^2 + dX^3 + eX^4 \ldots$ etc.

"Normal" Equations for a Second Degree Parabola

$$\begin{aligned} \text{(I)}\quad & \Sigma(Y) = Na + b\Sigma(X) + c\Sigma(X^2) \\ \text{(II)}\quad & \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2) + c\Sigma(X^3) \\ \text{(III)}\quad & \Sigma(X^2Y) = a\Sigma(X^2) + b\Sigma(X^3) + c\Sigma(X^4). \end{aligned} \qquad 65$$

Exponential Equations

$$\begin{aligned} Y &= a\,b^x \\ \log Y &= \log a + X \log b \\ Y &= a\,X^b \\ \log Y &= \log a + b \log X \end{aligned} \qquad 68, 93$$

Other Non Linear Curves

Hyperbola

$$Y = \frac{1}{a + bX} \qquad 95$$

Gompertz Curve

$$Y = a\,b^{c^x}$$

$$\log Y = \log a + C^x \log b \qquad 68$$

Pearl-Reed Curve

$$Y = \frac{k}{1 + e^{a + bX}} \qquad \text{Pearl}$$

Correlation

Coefficient of Correlation

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}} \qquad 83$$

Product moment method

$$r = \frac{p}{\sigma_x \sigma_y}$$

$$p = \frac{\Sigma(XY)}{N} - \left(\frac{\Sigma(X)}{N}\right)\left(\frac{\Sigma(Y)}{N}\right) \qquad 84$$

Standard Error of Estimate

$$S_y = \sqrt{\frac{\Sigma(d^2)}{N}} \qquad 79$$

$$S_y = \sigma_y \sqrt{1 - r^2}$$

Line of regression

$$y = r\frac{\sigma_y}{\sigma_x} x \qquad 84$$

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - X)$$

Coefficient of Determination

$$r^2 \qquad 88$$

Coefficient of Alienation

$$\sqrt{\frac{S_y^2}{\sigma_y^2}} \qquad 88$$

Coefficient of Non-Determination

$$k = \sqrt{1 - r^2} \qquad 88$$

Correction of S_y and r for number of Cases

$$\bar{S}_y^2 = S_y^2 \frac{(N-1)}{(N-2)} \qquad 88$$

$$\bar{r}^2 = 1 - (1 - r^2)\frac{(N-1)}{(N-2)} \qquad 88$$

Correlation from Ranks

$$\rho = 1 - \frac{6\,\Sigma(D^2)}{N\,(N^2 - 1)} \qquad 88$$

Relation Between Coefficient of Correlation From Ranks (ρ) and r

$$r = 2 \sin\left(\frac{\pi}{6}\,\rho\right) \qquad 90$$

Spearman's "Footrule"

$$R = 1 - \frac{6\Sigma G}{N^2 - 1} \qquad 90$$

Index of Correlation

$$\rho = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}} \qquad 95$$

$$\rho^2 = \frac{a\Sigma(Y) + b\Sigma(X\,Y) + c\Sigma(X^2\,Y) + \ldots - Nc_y^2}{\Sigma\,(Y^2) - Nc_y^2} \qquad 95$$

Correlation Ratio

$$\eta = \sqrt{1 - \frac{\sigma_{ay}^2}{\sigma_y^2}} \qquad 96$$

Correction of Correlation Ratio for Grouping

$$\eta'^2 = \frac{\eta^2 - \dfrac{(\kappa - 3)}{N}}{1 - \dfrac{(\kappa - 3)}{N}} \qquad 96$$

Test for Linearity of Regression

$$\zeta = \eta^2 - r^2 \qquad 97$$

Coefficient of Multiple Correlation

$$R_{1\cdot234} = \sqrt{1 - \frac{S^2_{1\cdot234}}{\sigma^2_1}} \qquad 99$$

$$R^2_{1\cdot234} = \frac{b_{12\cdot34}\,p_{12} + b_{13\cdot24}\,p_{13} + b_{14\cdot23}\,p_{14}}{\sigma^2_1} \qquad 99$$

Multiple Correlation Regression

Linear

$$X_1 = a + b_{12\cdot34}\,X_2 + b_{13\cdot24}\,X_3 + b_{14\cdot23}\,X_4 \qquad 98$$

"Normal" Equations for Multiple Correlation Regression

$$\text{(I)} \quad \Sigma(X_1) = Na + b_{12\cdot34}\Sigma(X_2) + b_{13\cdot24}\Sigma(X_3) + b_{14\cdot23}\Sigma(X_4)$$

$$\text{(II)} \quad \Sigma(X_1X_2) = a\Sigma(X_2) + b_{12\cdot34}\Sigma(X_2^2) + b_{13\cdot24}\Sigma(X_2X_3) + b_{14\cdot23}\Sigma(X_2X_4) \qquad 99$$

$$\text{(III)} \quad \Sigma(X_1X_3) = a\Sigma(X_3) + b_{12\cdot34}\Sigma(X_2X_3) + b_{13\cdot24}\Sigma(X_3^2) + b_{14\cdot23}\Sigma(X_3X_4)$$

$$\text{(IV)} \quad \Sigma(X_1X_4) = a\Sigma(X_4) + b_{12\cdot34}\Sigma(X_2X_4) + b_{13\cdot24}\Sigma(X_3X_4) + b_{14\cdot23}\Sigma(X_4^2)$$

$$\text{(I)} \quad p_{12} = b_{12\cdot34}\sigma^2_2 + b_{13\cdot24}p_{23} + b_{14\cdot23}p_{24}$$

$$\text{(II)} \quad p_{13} = b_{12\cdot34}p_{23} + b_{13\cdot24}\sigma^2_3 + b_{14\cdot23}p_{34} \qquad 99$$

$$\text{(III)} \quad p_{14} = b_{12\cdot34}p_{24} + b_{13\cdot24}p_{34} + b_{14\cdot23}\sigma^2_4$$

"a" may be obtained from the first "normal" equation

$$\Sigma(X) = Na + b_{12\cdot34}\Sigma(X_2) + b_{13\cdot24}\Sigma(X_3) + b_{14\cdot23}\Sigma(X_4)$$

Standard Error of Estimate for Multiple Correlation

$$S_{1\cdot234} = \sqrt{\frac{\Sigma d^2}{N}} \qquad 99$$

$$S^2_{1\cdot234} = \sigma^2_1 - b_{12\cdot34}p_{12} - b_{13\cdot24}p_{13} - b_{14\cdot23}p_{14} \qquad 99$$

Non Linear Multiple Correlation Regression

$$X_1 = a + f(X_2) + f(X_3) + f(X_4) + \text{etc.} \qquad 99$$

Coefficient of Partial Correlation

$$r_{13\cdot24} = \sqrt{b_{13\cdot24}\cdot b_{31\cdot24}} \qquad 100$$

$$r_{12\cdot3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r^2_{13}}\sqrt{1 - r^2_{23}}} \qquad 100$$

$$r_{14\cdot23} = 1 - \frac{(1 - R^2_{1\cdot234})}{(1 - R^2_{1\cdot23})} \qquad 100$$

Coefficient of Part Correlation

$${}_{12}r^2_{34} = \frac{b^2_{12\cdot34}\,\sigma_2{}^2}{b^2_{12\cdot34}\sigma_2{}^2 + \sigma_1{}^2(1 - R^2_{1\cdot234})} \qquad 100$$

Beta Coefficients

$$\beta_{12\cdot34} = b_{12\cdot34}\frac{\sigma_2}{\sigma_1}$$

$$\beta_{13\cdot24} = b_{13\cdot24}\frac{\sigma_3}{\sigma_1} \qquad \text{Ezekiel}$$

$$\beta_{14\cdot23} = b_{14\cdot23}\frac{\sigma_4}{\sigma_1}$$

Mean Square Contingency

$$\phi^2 = \frac{\chi^2}{N} \qquad 103$$

Coefficient of Contingency

$$CC = \sqrt{\frac{\chi^2}{N+\chi^2}} \qquad 104$$

$$CC = \sqrt{\frac{\phi^2}{1+\phi^2}} \qquad 104$$

Probability

Probability of Success

$$p = \frac{a}{N} \qquad 107$$

Probability of Failure

$$q = \frac{b}{N} \qquad 107$$

Arithmetic Mean of Bernoulli Distribution

$$\bar{X} = Np \qquad 108$$

Standard Deviation of Bernoulli Distribution

$$\sigma_B = \sqrt{Npq} \qquad 108$$

In Relative Form

$$\sigma_{B\%} = \sqrt{\frac{p\,q}{N}}$$

Standard Deviation of a Poisson Distribution

$$\sigma_P^2 = n\,p\,q - \Sigma(p_n - p)^2 \qquad \text{Rietz}$$

Standard Deviation of a Lexis Distribution

$$\sigma_L^2 = n\,p\,q + (n^2 - n)\,\sigma_{p_n}^2 \qquad \text{Rietz}$$

where

σ_{p_n} = standard deviation of the several probabilities $p_1, p_2, p_3 \ldots p_n$

Lexis Ratio

$$L = \frac{\sigma}{\sigma_B} \qquad \text{Rietz}$$

Charlier Coefficient of Disturbancy

$$100\rho = 100\,\frac{\sqrt{\sigma^2 - \sigma_B^2}}{n\,p} \qquad \text{Rietz}$$

Normal Curve

$$Y = Y_0 e^{\frac{-x^2}{2\sigma^2}} \qquad 112$$

$$y = \frac{N}{\sigma\sqrt{2\pi}}\, e^{\frac{-x^2}{2\sigma^2}} \qquad 113$$

Maximum Ordinate of Normal Curve

$$Y_0 = \frac{N}{\sigma\sqrt{2\pi}} = \frac{N}{2.506628\ \sigma} \qquad 113$$

Chi Square Test for Goodness of Fit

$$\chi^2 = \Sigma\left(\frac{(f_o - f)^2}{f}\right) \qquad 113$$

Theory of Sampling

Standard Error of Arithmetic Mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \qquad 120, 122$$

Probable Error of Arithmetic Mean

$$\text{P.E.}_{\bar{x}} = .6745\,\frac{\sigma}{\sqrt{N}} \qquad 120, 122$$

Standard Error of Median

$$\sigma_{mdn} = 1.2533\,\frac{\sigma}{\sqrt{N}} \qquad 122$$

Probable Error of Median

$$\text{P.E.}_{mdn} = .84535\,\frac{\sigma}{\sqrt{N}} \qquad 122$$

Standard Error and Probable Error of Standard Deviation for a sample drawn from a normally distributed universe.

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}}$$

$$\text{P.E.}_\sigma = .6745\,\frac{\sigma}{\sqrt{2N}} \qquad 122$$

Standard Error of Standard Deviation for a sample from any universe, normal or non-normal

$$\sigma_\sigma = \sqrt{\frac{\mu_4 - \mu_2^2}{4\,\mu_2 \cdot \mu}}$$

Standard Error and Probable Error of Mean Deviation

$$\sigma_{M.D.} = .6028\,\frac{\sigma}{\sqrt{N}}$$

$$\text{P.E.}_{M.D.} = .4066\,\frac{\sigma}{\sqrt{N}} \qquad 122$$

Standard Error and Probable Error of Coefficient of Variation

$$\sigma_v = \frac{V}{\sqrt{2N}}\sqrt{1 + \frac{2(V)^2}{(10)^4}} \qquad 122$$

$$\text{P.E.}_v = .6745\,\frac{V}{\sqrt{2N}}\sqrt{1 + \frac{2(V)^2}{(10)^4}}$$

Standard Error and Probable Error of Coefficient of Correlation

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$$

$$\text{P.E.}_r = .6745\,\frac{1 - r^2}{\sqrt{N}} \qquad 122, 129$$

Standard Error of Coefficient of Rank Correlation (ρ)

$$\sigma_\rho = \frac{1}{\sqrt{N - 1}}$$

Standard Error of Coefficient of Rank Correlation (Spearman's)

$$\sigma_\rho = \frac{1 - \rho^2}{\sqrt{N}}\,(1 + .086\rho^2 + .013\rho^4 + .002\rho^6) \qquad 122$$

$$\text{P.E.}_\rho = .6745\,\frac{1 - \rho^2}{\sqrt{N}}\,(1 + .086\rho^2 + .013\rho^4 + .002\rho^6)$$

Standard Error and Probable Error of Coefficient of Multiple Correlation

$$\sigma_{R1\cdot23\cdots n} = \frac{1 - R^2_{1\cdot23\cdots n}}{\sqrt{N}} \qquad 122$$

$$\text{P.E.}_{R1\cdot23\cdots n} = .6745\,\frac{1 - R^2_{1\cdot23\cdots n}}{\sqrt{N}}$$

Standard Error and Probable Error of Coefficient of Partial Correlation

$$\sigma_{r_{12\cdot 34\cdots n}} = \frac{1 - r^2_{12\cdot 34\cdots n}}{\sqrt{N}}$$

$$\text{P.E.}_{r_{12\cdot 34\cdots n}} = .6745\,\frac{1 - r^2_{12\cdot 34\cdots n}}{\sqrt{N}} \qquad 122$$

Standard Error of the Difference Between Two Means

$$\sigma_D = \sqrt{\sigma^2_{\bar{x}1} + \sigma^2_{\bar{x}2}} \qquad 124$$

$$= \sqrt{\frac{\sigma^2_1}{N_1} + \frac{\sigma^2_2}{N_2}} \qquad 124$$

Test for Significance of Coefficient of Correlation

$$z = \tfrac{1}{2}\left[(\log_e (1 + r) - \log_e (1 - r)\right] \qquad 129$$

$$\sigma_z = \frac{1}{\sqrt{N - 3}}$$

Standard Error of Second Moment About Mean

$$\sigma_{\mu_2} = \sqrt{\frac{\mu_4 - \mu^2_2}{N}}$$

$$\sigma_{\mu_2} = \sigma^2\sqrt{\frac{2}{N}}$$

Standard Error of Third Moment About Mean

$$\sigma_{\mu_3} = \sqrt{\frac{\mu_6 - \mu_3^2}{N}}$$

$$\sigma_{\mu_3} = \sigma_3\sqrt{\frac{6}{N}}$$

Standard Error of Fourth Moment About Mean

$$\sigma_{\mu_4} = \sqrt{\frac{\mu_8 - \mu_4^2}{N}}$$

$$\sigma_{\mu_4} = \sigma^4\sqrt{\frac{96}{N}}$$

Standard Error of β_2

$$\sigma_{\beta_2} = \sqrt{\frac{24}{N}}$$

Standard Error of Coefficient of Skewness as Measured by $\frac{\bar{X} - Mode}{\sigma}$

$$\sigma_{SK} = \sqrt{\frac{3}{2N}}$$

Standard Error of Semi-Interquartile Range

$$\sigma_Q = .7867 \frac{\sigma}{\sqrt{N}}$$

Standard Error of Difference Between Two Standard Deviations*

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\sigma^2_{\sigma_1} + \sigma^2_{\sigma_2}}$$

Standard Error of Difference Between Two Coefficients of Correlation*

$$\sigma_{r_{12} - r_{34}} = \sqrt{\sigma_{r_{12}}{}^2 - \sigma_{r_{34}}{}^2}$$

Standard Error of Coefficient of Regression

$$\sigma_{b_{xy}} = \frac{\sigma_x}{\sigma_y} \sqrt{\frac{1 - r^2_{xy}}{N}}$$

Standard Error of Correlation Ratio

$$\sigma_\eta = \frac{1 - \eta^2}{\sqrt{N}}$$

Standard Error of Arithmetic Mean for Small Samples

$$S_{\bar{x}} = \frac{s}{\sqrt{N}} \qquad 126$$

where $$s^2 = \frac{\Sigma(x^2)}{N - 1} = \frac{N\sigma^2}{N - 1}$$

Standard Error of Difference Between Two Arithmetic Means for Small Samples

$$S_D = \frac{s}{\sqrt{\frac{N_1 N_2}{N_1 + N_2}}} \qquad 127$$

$$s^2 = \frac{\Sigma(x_1^2) + \Sigma(x_2^2)}{N_1 + N_2 - 2}$$

*Mutually independent.

Standard Error of the Difference Between Two Proportions

$$\sigma_{D_{\%}} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \qquad 125$$

Standard Error of the Sum of a Series of Means with Given Standard Errors (when samples used are mutually independent)

$$\sigma^2_{\bar{X}1+\bar{X}2+\cdots+\bar{X}n} = \sigma_{\bar{X}1}^2 + \sigma_{\bar{X}2}^2 + \ldots + \sigma_{\bar{X}n}^2 \qquad 126$$

Standard Error of a Mean Raised to a Power

$$\frac{\sigma_{\bar{X}^n}}{\bar{\bar{X}}^n} = N\frac{\sigma_{\bar{X}}}{\bar{\bar{X}}} \qquad 127$$

Standard Error of a Product of a Series of Means (mutually independent)

$$\left(\frac{\sigma_{\bar{X}1\,\cdot\,\bar{X}2\,\cdots\,\bar{X}n}}{\bar{\bar{X}}_1 \cdot \bar{\bar{X}}_2 \cdots \bar{\bar{X}}_n}\right)^2 = \left(\frac{\sigma_{\bar{X}1}}{\bar{\bar{X}}_1}\right)^2 + \left(\frac{\sigma_{\bar{X}2}}{\bar{\bar{X}}_2}\right)^2 = \left(\frac{\sigma_{\bar{X}3}}{\bar{\bar{X}}_3}\right)^2 \cdots \left(\frac{\sigma_{\bar{X}n}}{\bar{\bar{X}}_n}\right)^2 \qquad 127$$

Standard Error of a Quotient of Two Means (mutually independent)

$$\left(\frac{\sigma_{\frac{\bar{X}1}{\bar{X}2}}}{\frac{\bar{\bar{X}}_1}{\bar{\bar{X}}_2}}\right) = \left(\frac{\sigma_{X1}}{\bar{\bar{X}}_1}\right)^2 + \left(\frac{\sigma_{\bar{X}2}}{\bar{\bar{X}}_2}\right)^2 \qquad 127$$

Standard Error of a Sum (Not mutually independent)

$$\sigma_{A+B} = \sqrt{\sigma_A^2 + \sigma_B^2 + 2r_{AB}\,\sigma_A\,\sigma_B}$$

Standard Error of a Difference (Not mutually independent)

$$\sigma_{A-B} = \sqrt{\sigma_A^2 + \sigma_B^2 - 2r_{AB}\,\sigma_A\,\sigma_B}$$

Index Numbers

Simple Aggregate of Actual Prices

$$\frac{\Sigma\, p_n}{\Sigma\, p_o} \qquad 134$$

Average of Relative Prices (Arithmetic Mean)

$$\frac{\Sigma\left(\frac{p_n}{p_o}\right)}{N} \qquad 136$$

Weighted Aggregate of Actual Prices

a. Base Year Weights

$$\frac{\Sigma\ (p_n q_o)}{\Sigma\ (p_o q_o)} \qquad 138$$

b. Given Year Weights

$$\frac{\Sigma\ (p_n q_n)}{\Sigma\ (p_o q_n)} \qquad 138$$

Weighted Average of Relative Prices

a. Arithmetic Mean (Base Year Weights)

$$\frac{\Sigma \left[\frac{p_n}{p_o} \times (p_o q_o) \right]}{\Sigma\ (p_o q_o)} \qquad 142$$

Given Year Weights

$$\frac{\Sigma \left[\frac{p_n}{p_o} \times (p_n q_n) \right]}{\Sigma\ (p_n q_n)} \qquad 142$$

b. Geometric Mean (Logarithmic Form—given year weights)

$$\text{Log Index} = \frac{\log \left[\frac{p'_n}{p'_o} (p'n\ q'n) \right] + \log \left[\frac{p''_n}{p''_o} (pn\ qn) \right] + \text{etc}}{\Sigma (pn\ qn)}$$

Ideal Index Number Formula

$$\sqrt{\frac{\Sigma\ (p_n q_o)}{\Sigma\ (p_o q_o)} \times \frac{\Sigma\ (p_n q_n)}{\Sigma\ (p_o q_n)}} \qquad 143$$

Weighted Average

Arithmetic Mean

$$\frac{\Sigma(\text{Items} \times \text{Weights})}{\Sigma(\text{weights})} = \frac{\Sigma(I \times W)}{\Sigma(W)} \qquad 137$$

Geometric Mean

$$\sqrt[\Sigma w]{I_1^{w} \cdot I_2^{w} \cdot I_3^{w} \cdots I_n^{w}}$$

Logarithm of Weighted Geometric Mean

$$\log \text{Weighted } G_n \qquad \frac{\Sigma(w \log I)}{\Sigma(w)}$$

Education and Psychology

Standard Score

$$z = \frac{x}{\sigma} \qquad 175$$

Conversion of Standard Scores (Average Score of 50)

$$z = 50 + \frac{x}{\sigma} 10 \qquad 175$$

Coefficient of Correlation from Standard Scores

$$r_{12} = \frac{\Sigma(z_1 z_2)}{N} \qquad \text{Kelley}$$

Coefficient of Reliability Resulting From Test Given n Times

$$r_n = \frac{n\, r_{11}}{1 + (n-1)\, r_{11}} \qquad 176$$

Intelligence Quotient

$$\text{I. Q.} = \frac{\text{M. A.}}{\text{C. A.}} \qquad 176$$

Subject Quotients

Arithmetic Quotient

$$\frac{\text{Arithmetic Age}}{\text{Chronological Age}} \qquad 177$$

Reading Quotient

$$\frac{\text{Reading Age}}{\text{Chronological Age}} \qquad 177$$

Any Subject Quotient

$$\frac{\text{Any Subject Age}}{\text{Chronological Age}} \qquad 177$$

Subject Ratio

$$\frac{\text{Subject Age}}{\text{Mental Age}} \qquad 177$$

Biology

Index of Abmodality

$$\frac{x}{\sigma} \qquad 177$$

Variability of Offspring

$$\sigma_{3 \cdot 12} = \sigma_3 \sqrt{1 - \frac{2r^2_1}{1 + r_3}} \qquad 178$$

assuming $r_1 = r_2$

Abmodality of Offspring

$$h_3 = \frac{r_1 \sigma_3}{(1 + r_3)\sigma_1} \left(h_1 + \frac{\sigma_1}{\sigma_2} h_2\right) \qquad 178$$

assuming $r_1 = r_2$

Miscellaneous Formulas

Sum of Squares of First n natural numbers

$$\Sigma(n^2) = \frac{2n^3 + 3n^2 + n}{6} \qquad \text{Camp}$$

Number of Combinations of n Things Taken r at a Time

$$nCr = \frac{n(n-1)(n-2)\ldots\ldots(n-r+1)}{r(r-1)(r-2)\ldots\ldots 1} = \frac{nPr}{r!} \qquad \text{Camp}$$

Number of Permutations of n Things Taken r at a Time

$$nPr = \frac{n!}{(n-r)!} \qquad \text{Camp}$$

REFERENCES

CAMP, B. H., *The Mathematical Part of Elementary Statistics.* D. C. Heath and Company, Boston, 1931.

KELLEY, TRUMEN L., *Statistical Method.* MacMillan Co., New York, 1925.

PEARL, RAYMOND, *Medical Biometry and Statistics.* W. B. Saunders, Philadelphia, 1930.

REITZ, H. L., (Editor), *Handbook of Mathematical Statistics.* Houghton, Mifflin Co., New York, 1924.

LIST OF SYMBOLS

a	— Y intercept
a	— number of ways in which a favorable outcome can appear
b	— coefficient of slope (or regression)
b	— possible number of unfavorable results
$b_{12.34}$	— coefficient of net regression of X_2 on X_1 excluding X_3 and X_4
C	— size of class interval
CA	— chronological age
CC	— coefficient of contingency
c	— difference between arbitrary origin and mean or median
c	— constant in trend or regression equation
D	— differences in rank
d	— deviation of an individual value or *midpoint of a* class interval from an arbitrary or guessed mean (or average—other than arithmetic mean)
d	— deviation from line of trend $(Y - Yc)$
d	— constant in trend or regression equation
d'	— d in class interval units
e	— a constant = 2.71828
f	— frequency
f_a	— frequency of class interval above modal group
f_b	— frequency of class interval below modal group
G	— Positive differences in rank
G_m	— Geometric Mean
H_m	— Harmonic Mean
h_1	— deviation (abmodality) of father
h_2	— deviation (abmodality) of mother
h_3	— deviation of mean of given characteristic of offspring from mean of characteristic of all offspring
I	— item
I.Q.	— intelligence quotient
k	— coefficient of non-determination
L_{md}	— lower limit of class interval containing median
L_{mo}	— lower limit of modal group
$M.A.$	— mental age

MD — mean deviation
MD' — mean deviation in class intervals
$M.P.$ — mid point
N or n — number of cases
N_L — number of cases overstated or too large
N_S — number of cases understated or too small
$P.E._{\bar{x}}$ — probable error of mean
$P.E._{\theta}$ — probable error of any statistical measure (θ) such as $P.E._{mdn}$, $P.E._{r}$, etc.
p — product moment
p — probability of success
p_n — price of a commodity in period n
p_o — price of a commodity in base period
p_o' — price of first commodity in base period
p_o'' — price of second commodity in base period
p_1 — price of a commodity in first period
p_2 — price of a commodity in second period
Q_1 — first quartile
Q_3 — third quartile
QD — quartile deviation (semi-inter-quartile range)
Q_m — quadratic mean
q — probability of failure
q_n — quantity of a commodity produced or consumed in period n
q_o — quantity of a commodity produced or consumed in base period
q_o' — quantity of first commodity produced or consumed in base period
q_o'' — quantity of second commodity produced or consumed in base period
q_1 — quantity of a commodity produced or consumed in first period
q_2 — quantity of a commodity produced or consumed in second period
R — coefficient of correlation computed by Spearman's "footrule" method
$R_{1.234}$ — coefficient of multiple correlation between X_1 and X_2, X_3, X_4
r or r_{12} or r_{xy} — coefficient of correlation
$\bar{r}$ — coefficient of correlation corrected for number of cases
r_1 — coefficient of heredity (fathers and off-spring)

r_2 — coefficient of heredity (mothers and off-spring)

r_3 — coefficient of assortative mating (fathers and mothers)

r_{11} — coefficient of reliability

$r_{13.2}$ — coefficient of partial correlation between X_1 and X_3 excluding X_2

$_{12}r_{34}$ — coefficient of part correlation between X_1 and X_3 excluding X_2 and X_4

$S_{1.234}$ — standard error of estimate measured about regression surface $X_1 = f(X_2) + f(X_3) + f(X_4)$

S_y — standard error of estimate

$\bar{S}_y$ — standard error of estimate corrected for number of cases

W or Wt — weight

X — an individual value

$\overline{X}$ — arithmetic mean

x — deviation of individual value from its arithmetic mean

Y — an individual value

$\overline{Y}$ — arithmetic mean of Y values

Y_c — computed Y value as determined from line of trend or regression

y — deviation of individual Y value from its arithmetic mean

$\overline{Z}$ — an arbitrarily selected value-guessed mean

z — standard score

z' — residual-difference between actual value and theoretical line of regression value

z_1 — standard score on first test

z_2 — standard score on second test

α_3 — measure of skewness

β_1 — curve criterion

β_2 — curve criterion (measure of kurtosis)

$\beta_{12.3}$, $\beta_{12.34}$, $\beta_{13.24}$, $\beta_{24.23}$ — beta coefficients

ζ — test for linearity of regression

η — correlation ratio

θ — any statistic

κ — curve criterion

κ — number of arrays in correlation table

μ_1' μ_2' μ_3' μ_4' — moments about arithmetic mean corrected for grouping (Sheppard's Corrections)

μ_2 μ_3 μ_4 — moments about arithmetic mean

ν_1 ν_2 ν_3 — moments about arbitrary origin

π — a constant = 3.141593

ρ — index of correlation (measured on basis of curvilinear line of regression)

ρ — coefficient of correlation from Ranks

Σ — sum of

σ — standard deviation

σ' — standard deviation corrected for grouping error

σ_{xy} — standard deviation of values about means of respective columns in correlation table

σ_λ — standard error of arithmetic mean, the standard error of any statistical measure (σ_{mdn}, σ_r, σ_s, etc.) is similarly written

$\sigma_{3 \cdot 12}$ — standard deviation of an array of off-spring

σ_s — standard deviation of off-spring in general

ϕ^2 — mean squared contingency

χ — measure of skewness

χ^2 — chi square, value used in test for goodness of fit

Table of Logarithms

	0	1	2	3	4	5	6	7	8	9
10	.00000	.00432	.00860	.01284	.01703	.02119	.02531	.02938	.03342	.03743
11	.04139	.04532	.04922	.05308	.05690	.06070	.06446	.06819	.07188	.07555
12	.07918	.08279	.08636	.08991	.09342	.09691	.10037	.10380	.10721	.11059
13	.11394	.11727	.12057	.12385	.12710	.13033	.13354	.13672	.13988	.14301
14	.14613	.14922	.15229	.15534	.15836	.16137	.16435	.16732	.17026	.17319
15	.17609	.17898	.18184	.18469	.18752	.19033	.19312	.19590	.19866	.20140
16	.20412	.20683	.20952	.21219	.21484	.21748	.22011	.22272	.22531	.22789
17	.23045	.23300	.23553	.23805	.24055	.24304	.24551	.24797	.25042	.25285
18	.25527	.25768	.26007	.26245	.26482	.26717	.26951	.27184	.27416	.27646
19	.27875	.28103	.28330	.28556	.28780	.29003	.29226	.29447	.29667	.29885
20	.30103	.30320	.30535	.30750	.30963	.31175	.31387	.31597	.31806	.32015
21	.32222	.32428	.32634	.32838	.33041	.33244	.33445	.33646	.33846	.34044
22	.34242	.34439	.34635	.34830	.35025	.35218	.35411	.35603	.35793	.35984
23	.36173	.36361	.36549	.36736	.36922	.37107	.37291	.37475	.37658	.37840
24	.38021	.38202	.38382	.38561	.38739	.38917	.39094	.39270	.39445	.39620
25	.39794	.39967	.40140	.40312	.40483	.40654	.40824	.40993	.41162	.41330
26	.41497	.41664	.41830	.41996	.42160	.42325	.42488	.42651	.42813	.42975
27	.43136	.43297	.43457	.43616	.43775	.43933	.44091	.44248	.44404	.44560
28	.44716	.44871	.45025	.45179	.45332	.45484	.45637	.45788	.45939	.46090
29	.46240	.46389	.46538	.46687	.46835	.46982	.47129	.47276	.47422	.47567
30	.47712	.47857	.48001	.48144	.48287	.48430	.48572	.48714	.48855	.48996
31	.49136	.49276	.49415	.49554	.49693	.49831	.49969	.50106	.50243	.50379
32	.50515	.50651	.50786	.50920	.51055	.51188	.51322	.51455	.51587	.51720
33	.51851	.51983	.52114	.52244	.52375	.52504	.52634	.52763	.52892	.53020
34	.53148	.53275	.53403	.53529	.53656	.53782	.53908	.54033	.54158	.54283
35	.54407	.54531	.54654	.54777	.54900	.55023	.55145	.55267	.55388	.55509
36	.55630	.55751	.55871	.55991	.56110	.56229	.56348	.56467	.56585	.56703
37	.56820	.56937	.57054	.57171	.57287	.57403	.57519	.57634	.57749	.57864
38	.57978	.58092	.58206	.58320	.58433	.58546	.58659	.58771	.58883	.58995
39	.59106	.59218	.59329	.59439	.59550	.59660	.59770	.59879	.59988	.60097
40	.60206	.60314	.60423	.60531	.60638	.60746	.60853	.60959	.61066	.61172
41	.61278	.61384	.61490	.61595	.61700	.61805	.61909	.62014	.62118	.62221
42	.62325	.62428	.62531	.62634	.62737	.62839	.62941	.63043	.63144	.63246
43	.63347	.63448	.63548	.63649	.63749	.63849	.63949	.64048	.64147	.64246
44	.64345	.64444	.64542	.64640	.64738	.64836	.64933	.65031	.65128	.65225
45	.65321	.65418	.65514	.65610	.65706	.65801	.65896	.65992	.66087	.66181
46	.66276	.66370	.66464	.66558	.66652	.66745	.66839	.66932	.67025	.67117
47	.67210	.67302	.67394	.67486	.67578	.67669	.67761	.67852	.67943	.68034
48	.68124	.68215	.68305	.68395	.68485	.68574	.68664	.68753	.68842	.68931
49	.69020	.69108	.69197	.69285	.69373	.69461	.69548	.69636	.69723	.69810
50	.69897	.69984	.70070	.70157	.70243	.70329	.70415	.70501	.70586	.70672
51	.70757	.70842	.70927	.71012	.71096	.71181	.71265	.71349	.71433	.71517
52	.71600	.71684	.71767	.71850	.71933	.72016	.72099	.72181	.72263	.72346
53	.72428	.72509	.72591	.72673	.72754	.72835	.72916	.72997	.73078	.73159
54	.73239	.73320	.73400	.73480	.73560	.73640	.73719	.73799	.73878	.73957
55	.74036	.74115	.74194	.74273	.74351	.74429	.74507	.74586	.74663	.74741
56	.74819	.74896	.74974	.75051	.75128	.75205	.75282	.75358	.75435	.75511
57	.75587	.75664	.75740	.75815	.75891	.75967	.76042	.76118	.76193	.76268
58	.76343	.76418	.76492	.76567	.76641	.76716	.76790	.76864	.76938	.77012
59	.77085	.77159	.77232	.77305	.77379	.77452	.77525	.77597	.77670	.77743

Table of Logarithms (continued)

	0	1	2	3	4	5	6	7	8	9
60	77815	77887	.77960	.78032	.78104	.78176	78247	78319	.78390	.78462
61	78533	78604	78675	.78746	.78817	.78888	.78958	.79029	.79099	.79169
62	79239	79309	.79379	.79449	.79518	.79588	.79657	.79727	.79796	.79865
63	79934	80003	.80072	80140	.80209	.80277	80346	.80414	.80482	.80550
64	80618	80686	.80754	.80821	.80889	.80956	81023	.81090	.81158	.81224
65	81291	.81358	.81425	.81491	.81558	.81624	.81690	.81757	.81823	.81889
66	81954	82020	.82086	.82151	.82217	.82282	82347	.82413	.82478	82543
67	82607	82672	82737	.82802	.82866	82930	82995	83059	83123	83187
68	.83251	.83315	83378	.83442	.83506	.83569	83632	.83696	.83759	.83822
69	83885	83948	.84011	.84073	.84136	84198	.84261	.84323	.84386	84448
70	.84510	84572	.84634	.84696	.84757	.84819	.84880	.84942	.85003	.85065
71	.85126	85187	.85248	.85309	.85370	.85431	.85491	.85552	85612	.85673
72	.85733	85794	.85854	.85914	.85974	.86034	.86094	.86153	.86213	.86273
73	86332	.86392	.86451	.86510	.86570	.86629	.86688	.86747	.86806	86864
74	.86923	.86982	.87040	.87099	.87157	.87216	.87274	.87332	.87390	87448
75	.87506	87564	.87622	.87679	.87737	.87795	.87852	87910	.87967	88024
76	.88081	88138	.88195	.88252	.88309	.88366	.88423	88480	.88536	88593
77	.88649	.88705	.88762	.88818	.88874	.88930	.88986	89042	.89098	89154
78	.89209	.89265	.89321	.89376	.89432	.89487	.89542	.89597	.89653	.89708
79	.89763	.89818	.89873	.89927	.89982	.90037	.90091	.90146	90200	.90255
80	.90309	90363	.90417	.90472	.90526	.90580	.90634	.90687	.90741	.90795
81	.90849	90902	.90956	.91009	.91062	.91116	.91169	.91222	.91275	.91328
82	.91381	.91434	.91487	.91540	.91593	.91645	.91698	.91751	.91803	.91855
83	.91908	.91960	.92012	.92065	.92117	.92169	.92221	.92273	.92324	.92376
84	.92428	.92480	.92531	.92583	.92634	.92686	.92737	.92788	.92840	.92891
85	.92942	.92993	.93044	.93095	.93146	.93197	.93247	.93298	.93349	.93399
86	.93450	.93500	.93551	.93601	.93651	.93702	93752	93802	.93852	.93902
87	93952	94002	.94052	.94101	.94151	.94201	.94250	.94300	.94349	.94399
88	.94448	.94498	.94547	.94596	.94645	.94694	.94743	.94792	.94841	.94890
89	.94939	.94988	.95036	.95085	.95134	.95182	.95231	.95279	.95328	.95376
90	95424	.95472	.95521	.95569	.95617	.95665	95713	.95761	.95809	.95856
91	95904	.95952	.95999	.96047	.96095	.96142	.96190	.96237	.96284	.96332
92	.96379	.96426	.96473	.96520	.96567	.96614	.96661	.96708	.96755	.96802
93	.96848	.96895	.96942	.96988	.97035	.97081	.97128	97174	.97220	.97267
94	.97313	.97359	.97405	.97451	.97497	.97543	.97589	.97635	.97681	.97727
95	.97772	.97818	.97864	.97909	.97955	.98000	.98046	.98091	.98137	.98182
96	.98227	.98272	.98318	.98363	.98408	.98453	.98498	98543	.98588	98632
97	.98677	.98722	.98767	.98811	.98856	.98900	.98945	.98989	.99034	.99078
98	.99123	99167	.99211	.99255	.99300	.99344	.99388	.99432	.99476	.99520
99	.99564	.99607	.99651	.99695	.99739	.99782	.99826	.99870	.99913	.99957

TECHNICAL APPENDIX I

DERIVATION OF SHORT METHOD OF COMPUTING ARITHMETIC MEAN [1]

For Ungrouped Data

Let each value (X) equal

$$X = \bar{Z} + d$$

or the arbitrary starting point plus the deviation of the value from that point.

The total of all values will then be

$$\Sigma\bar{Z} + \Sigma d$$

but since $\bar{Z}$ is a constant, this may be rewritten

$$\Sigma X = N\bar{Z} + \Sigma d$$

Dividing the total by N to obtain the arithmetic mean the result is

$$\frac{\Sigma X}{N} = \bar{Z} + \frac{\Sigma d}{N}$$

or

$$\bar{X} = \bar{Z} + \frac{\Sigma d}{N}$$

For Grouped Data

The midpoint of each group may be measured as a deviation from the guessed mean.

$$M.P. = \bar{Z} + d$$

To obtain the total value of all cases in the class interval the midpoint of each group is multiplied by the number of cases in the group

$$f \times MP = f\bar{Z} + fd$$

Totaling up for all class intervals

$$\Sigma(f \times MP) = \Sigma(f\bar{Z}) + \Sigma(fd)$$

or since $\bar{Z}$ is a constant

$$\Sigma(f \times MP) = \bar{Z}\Sigma(f) + \Sigma(fd)$$

and since $\Sigma(f) = N$

$$\Sigma(f \times MP) = N\bar{Z} + \Sigma(fd)$$

[1] This derivation is after Yule.

dividing by N to obtain the arithmetic mean

$$\frac{\Sigma(f \times MP)}{N} = \bar{Z} + \frac{\Sigma(fd)}{N}$$

or

$$\overline{X} = \bar{Z} + \frac{\Sigma(fd)}{N}$$

TECHNICAL APPENDIX II

DERIVATION OF SHORT FORMULA FOR STANDARD DEVIATION

If d, is the deviation of a given point from an arbitrary origin $(\bar{Z})$

$$d = X - \bar{Z}$$

and if the difference between the mean and this origin is termed c

$$c = \overline{X} - \bar{Z}$$

then

$$\begin{aligned} d - c &= (X - \bar{Z}) - (\overline{X} - \bar{Z}) \\ &= X - \bar{Z} - \overline{X} + \bar{Z} \\ &= X - \overline{X} \text{ or } x \end{aligned}$$

where x is the deviation of a value from the arithmetic mean but

$$\overline{X} = \bar{Z} + \frac{\Sigma(fd)}{N}$$

$$\therefore \quad \overline{X} - \bar{Z} = \frac{\Sigma(fd)}{N} = c$$

Since

$$\begin{aligned} d - c &= x \\ d &= x + c \\ \therefore \quad d^2 &= x^2 + 2cx + c^2 \end{aligned}$$

and

$$f(d^2) = f(x^2) + 2cfx + fc^2$$

and

$$\begin{aligned} \Sigma f(d^2) &= \Sigma f(x^2) + 2c\Sigma(fx) + c^2\Sigma f \\ &= \Sigma f(x^2) + 2c\Sigma(fx) + Nc^2 \end{aligned}$$

For

$$\Sigma f = N$$

But the sum of the deviations about the arithmetic mean is zero

$$\therefore \quad \Sigma(fx) = 0$$

and the formula reduces to

$$\Sigma f(d^2) = \Sigma f(x^2) + Nc^2$$

or

$$\Sigma f(x^2) = \Sigma f(d^2) - Nc^2$$

and

$$\frac{\Sigma f(x^2)}{N} = \frac{\Sigma f(d^2)}{N} - c$$

but

$$\sigma = \sqrt{\frac{\Sigma f(x^2)}{N}} \qquad \text{(see page 36)}$$

$$\therefore \sigma = \sqrt{\frac{\Sigma f(d^2)}{N} - c^2}$$

but as above

$$c = \frac{\Sigma(fd)}{N}$$

and

$$\sigma = \sqrt{\frac{\Sigma f(d^2)}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

TECHNICAL APPENDIX III

DERIVATION OF SHORT FORMULA FOR STANDARD DEVIATION—UNGROUPED DATA

A simpler formula may be arrived at for ungrouped data by selecting zero as the arbitrary origin then

$$d = X - \bar{Z}$$

but since

$$\bar{Z} = 0$$

$$d = X$$

and

$$c = \bar{X} - \bar{Z}$$

$$\therefore c = \bar{X}$$

since

$$d^2 = x^2 + 2cx + c^2 \quad \text{(see page 212)}$$

and

$$d = X$$

$$\therefore X^2 = x^2 + 2cx + c^2$$

and

$$\Sigma(X^2) = \Sigma(x^2) + 2c\Sigma(x) + Nc^2$$

but

$$\Sigma(x) = 0 \text{ and } c = \bar{X}$$

$$\therefore \Sigma(X^2) = \Sigma(x^2) - N\bar{X}^2$$

and

$$\Sigma(x^2) = \Sigma(X^2) - N\bar{X}^2$$

$$\frac{\Sigma(x^2)}{N} = \frac{\Sigma(X^2)}{N} - (\bar{X}^2) = \frac{\Sigma(X^2)}{N} - \left(\frac{\Sigma X}{N}\right)^2$$

Since

$$\sigma = \sqrt{\frac{\Sigma(x^2)}{N}}$$

$$\therefore \sigma = \sqrt{\frac{\Sigma(X^2)}{N} - \left(\frac{\Sigma X}{N}\right)^2}$$

TECHNICAL APPENDIX IV

DERIVATION OF "NORMAL" EQUATIONS FOR LEAST SQUARES STRAIGHT LINE

As shown on pages 52-54 of the text the formula for any straight line will be

$$Y_c = a + bX$$

where Yc represents the computed or theoretical value for Y obtained by substituting the appropriate value in the formula.

The problem is to determine a line which will fulfill the conditions of the **principle of least squares;** i.e., the sums of the squares of the deviations of the actual from the theoretical values will be a minimum.

The letter d may be used to represent the difference between the actual and theoretical values. The purpose is then to obtain a line so that:

$$\Sigma(d^2) = \text{a minimum}$$

$$\text{but } d = Y - Y_c$$

$\Sigma(Y - Y_c)^2$ must equal a minimum. We may then obtain the partial derivatives with respect to a and b and equate to zero to obtain a minimum.

$$\frac{\partial(Y - Y_c)^2}{\partial a} = 2Na - 2\Sigma(Y) + 2b\Sigma(X)$$

and

$$\frac{\partial(Y - Y_c)^2}{\partial b} = 2b\Sigma(X^2) - 2\Sigma(YX) + 2a\Sigma(X)$$

Equating to zero:

$$\text{I} \quad 2Na - 2\Sigma(Y) + 2b\Sigma(X) = 0$$
$$\text{II} \quad 2b\Sigma(X^2) - 2\Sigma(XY) + 2a\Sigma(X) = 0$$

or

$$\text{I} \quad \Sigma(Y) = Na + b\Sigma(X)$$
$$\text{II} \quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2)$$

TECHNICAL APPENDIX V

DERIVATION OF PRODUCT MOMENT FORMULA FOR COEFFICIENT OF CORRELATION

The original formula for r is

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

where

$$S_y = \sqrt{\frac{\Sigma(d^2)}{N}} \tag{1}$$

$$S_y^2 = \frac{\Sigma(d^2)}{N} \tag{2}$$

assuming a straight line regression

$$Y_c = a + bX \tag{3}$$

Y_c is used for the theoretical value obtained from the equation but

$$d = Y - Y_c \tag{4}$$
$$\therefore d = Y - (a + bX) \tag{5}$$
$$d = Y - a - bX \tag{6}$$

multiplying by d

$$d^2 = dY - ad - bdX \tag{7}$$

since there is one d for each value a summation is made for a points

$$\Sigma(d^2) = \Sigma(dY) - a\Sigma(d) - b\Sigma(dX) \tag{8}$$

Since the regression line is fitted by the least squares method

$$\Sigma(d) = 0$$
$$\Sigma(dX) = 0$$
$$\therefore \Sigma(d^2) = \Sigma(dY) \tag{9}$$

multiplying

$$d = Y - a - bX$$

by Y and summing up

$$\Sigma(dY) = \Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) \tag{10}$$

but since

$$\Sigma(d^2) = \Sigma(dY)$$

$$\therefore \Sigma(d^2) = \Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) \tag{11}$$

and

$$\frac{\Sigma(d^2)}{N} = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY)}{N} \tag{12}$$

but

$$S_y^2 = \frac{\Sigma d^2}{N}$$

$$\therefore S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY)}{N} \tag{13}$$

and

$$\sigma_y = \sqrt{\frac{\Sigma(Y^2)}{N} - c_y^2} \tag{14}$$

while

$$\sigma_y^2 = \frac{\Sigma(Y^2)}{N} - c_y^2 \tag{15}$$

substituting

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

$$r^2 = 1 - \frac{\dfrac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY)}{N}}{\dfrac{\Sigma(Y^2)}{N} - c_y^2} \tag{16}$$

multiplying numerator and denominator of fraction by N

$$r^2 = 1 - \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY)}{\Sigma(Y^2) - Nc_y^2} \tag{17}$$

This formula may be reduced to[1]

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2} \tag{18}$$

The two normal equations for the line of regression are

$$\text{(I)} \quad \Sigma(Y) = Na + b\Sigma(X) \tag{19}$$

$$\text{(III)} \quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2) \tag{20}$$

[1] This formula is known as the "least squares" formula for the coefficient of correlation.

If the point of averages ($\overline{X}$ and $\overline{Y}$) is used as an origin all values will be reduced to deviations from their respective means (x and y)

where

$$x = X - \overline{X}$$
$$y = Y - \overline{Y}$$

So that the equations will read

$$\text{(I)} \quad \Sigma(y) = Na + b\Sigma(x) \tag{21}$$
$$\text{(II)} \quad \Sigma(xy) = a\Sigma(x) + b\Sigma(x^2) \tag{22}$$

but since the sum of the deviations about the arithmetic mean equals zero

$$\therefore \Sigma(x) = 0$$
$$\Sigma(y) = 0$$

and the normal equation will reduce to

$$\text{(I)} \quad Na = 0 \tag{23}$$
$$\therefore a = 0$$
$$\text{(II)} \quad \Sigma(xy) = b\Sigma(x^2)$$
$$\therefore b = \frac{\Sigma(xy)}{\Sigma(x^2)} \tag{24}$$

reducing equation (17) into terms of deviations from the point of averages[1]

$$r^2 = \frac{a\Sigma(y) + b\Sigma(xy) - Nc_y^2}{\Sigma(y^2) - Nc_y^2} \tag{25}$$

but

$$\Sigma(y) = 0$$
$$a\Sigma(y) = 0$$
$$\text{and } c_y = 0$$

the equation thus reduces to

$$r^2 = \frac{b\Sigma(xy)}{\Sigma(y^2)} \tag{26}$$

but from (22)

$$b = \frac{\Sigma(xy)}{\Sigma(x^2)}$$

$$\therefore r^2 = \frac{\Sigma(xy)}{\Sigma(x^2)} \cdot \frac{\Sigma(xy)}{\Sigma(y^2)} \tag{27}$$

$$r^2 = \frac{[\Sigma(xy)]^2}{\Sigma(x^2) \cdot \Sigma(y^2)} \tag{28}$$

[1] Since "a" (the Y intercept) equals zero the line will pass through the origin or the point of averages

dividing numerator and denominator by N^2

$$r^2 = \frac{\left(\frac{\Sigma(xy)}{N}\right)^2}{\frac{\Sigma(x^2)}{N} \quad \frac{\Sigma(y^2)}{N}} \tag{29}$$

but

$$\sigma_x^2 = \frac{\Sigma(x^2)}{N} \quad \text{(see chapter IV)}$$

$$\sigma_y^2 = \frac{\Sigma(y^2)}{N} \quad \text{(see chapter IV)}$$

$$\therefore r = \frac{p}{\sigma_x \ \sigma_y} \tag{30}$$

where

$$p = \frac{\Sigma(xy)}{N} \tag{31}$$

Using the values X and Y as deviations from an arbitrary origin (in the case of ungrouped data, zero, so that the original values may be used) p may be computed from

$$p = \frac{\Sigma(xy)}{N} - c_x c_y \tag{32}$$

where x' and y' are deviations from arbitrary selected points for

$$x' = x + c_x$$

where c_x is the difference between the true mean and an arbitrary origin ($\frac{\Sigma(fd)}{N}$ for grouped data and $\frac{\Sigma(X)}{N}$ for ungrouped where zero is selected as an origin)

$$\begin{aligned} y' &= y + c_y \\ x'y' &= xy + c_x y + c_y x + c_x c_y \end{aligned} \tag{33}$$

summing up for all points

$$\Sigma(x'y') = \Sigma(xy) + c_x\Sigma(y) + c_y\Sigma(x) + Nc_x c_y \tag{34}$$

but since the sum of deviations about the means total up to zero

$$\Sigma(y) = 0$$
$$\Sigma(x) = 0$$

and equation 34 reduces to

$$\Sigma(x'y') = \Sigma(xy) + Nc_x c_y$$

dividing by N

$$\frac{\Sigma(x'y')}{N} = \frac{\Sigma(xy)}{N} + c_x c_y$$

$$\therefore \frac{\Sigma(xy)}{N} = \frac{\Sigma(x'y')}{N} - c_x c_y = p$$

If the arbitrary origin used is zero

$$x' = X$$
$$y' = Y$$

and

$$p = \frac{\Sigma(xy)}{N} = \frac{\Sigma(XY)}{N} - c_x c_y$$

TECHNICAL APPENDIX VI

DERIVATION OF FORMULA FOR LINE OF REGRESSION

Since the regression line is assumed to be straight its formula will be of the type

$$Y = a + bX$$

with the two "normal" equations[1]

$$\text{(I)} \quad \Sigma(Y) = Na + b\Sigma(X)$$
$$\text{(II)} \quad \Sigma(XY) = a\Sigma(X) + b\Sigma(X^2)$$

If the origin of the line is assumed to be at the point of averages the normal equations will read

$$\text{(I)} \quad \Sigma(y) = Na + b\Sigma(x)$$
$$\text{(II)} \quad \Sigma(xy) = \Sigma(x) = b\Sigma(x^2)$$

but

$$\Sigma(y) = 0$$
$$\Sigma(x) = 0$$
$$\therefore \text{(I)} \quad Na = 0 \text{ and } a = 0$$
$$\text{(II)} \quad \Sigma(xy) = b\Sigma(x^2) \text{ and } b = \frac{\Sigma(xy)}{\Sigma(x^2)}$$

Equation (1) will reduce to

$$y = bx$$

where

$$b = \frac{\Sigma(xy)}{\Sigma(x^2)}$$

(a) See chapter VI.

dividing numerator and denominator by N

$$y = \frac{\Sigma(xy)}{N\frac{\Sigma(x^2)}{N}}\,x$$

but

$$\frac{\Sigma(x^2)}{N} = \sigma^2_x$$

$$\therefore\ y = \frac{\Sigma(xy)}{N\sigma^2_x}\,x$$

but

$$\frac{\Sigma(xy)}{N\sigma^2_x} = \frac{\Sigma(xy)}{N\sigma_x\sigma_y}\quad\frac{\sigma_y}{\sigma_x}$$

and

$$r = \frac{\Sigma(xy)}{N\sigma_x\sigma_y}\ \text{(product moment formula)}$$

$$\therefore\ y = r\frac{\sigma_y}{\sigma_x}\,x$$

TECHNICAL APPENDIX VII

MULTIPLE CORRELATION REGRESSION

The "normal" equations for three independent variables, linear correlation with the type formula

$$X_1 = a + b_{12\cdot34}\,X_2 + b_{13\cdot24}\,X_3 + b_{14\cdot23}\,X_4$$

are:

(I) $\Sigma(X_1) = Na + b_{12\cdot34}\,\Sigma(X_2) + b_{13\cdot24}\,\Sigma(X_3) + b_{14\cdot23}\,\Sigma(X_4)$

(II) $\Sigma(X_1X_2) = a\Sigma(X_2) + b_{12\cdot34}\,\Sigma(X_2^2) + b_{13\cdot24}\,\Sigma(X_2X_3) + b_{14\cdot23}\,\Sigma(X_2X_4)$

(III) $\Sigma(X_1X_3) = a\Sigma(X_3) + b_{12\cdot34}\,\Sigma(X_2X_3) + b_{13\cdot24}\,\Sigma(X_3^2) + b_{14\cdot23}\,\Sigma(X_3X_4)$

(IV) $\Sigma(X_1X_4) = a\Sigma(X_4) + b_{12\cdot34}\,\Sigma(X_2X_4) + b_{13\cdot24}\,\Sigma(X_3X_4) + b_{14\cdot23}\,\Sigma(X_4^2)$

These equations may be simplified by assuming the origin to be at the point of averages and dividing both sides of the equations by N

(I) $\frac{\Sigma(x_1)}{N} = a + b_{12\cdot34}\frac{\Sigma(x_2)}{N} + b_{13\cdot24}\frac{\Sigma(x_3)}{N} + b_{14\cdot23}\frac{\Sigma(x_4)}{N}$

$$\text{(II)}\quad \frac{\Sigma(x_1x_2)}{N} = a\frac{\Sigma(x_2)}{N} + b_{12\cdot34}\frac{\Sigma(x_2^2)}{N} + b_{13\cdot24}\frac{\Sigma(x_2x_3)}{N} + b_{14\cdot23}\frac{\Sigma(x_2x_4)}{N}$$

$$\text{(III)}\quad \frac{\Sigma(x_1x_3)}{N} = a\frac{\Sigma(x_3)}{N} + b_{12\cdot34}\frac{\Sigma(x_2x_3)}{N} + b_{13\cdot24}\frac{\Sigma(x_3^2)}{N} + b_{14\cdot23}\frac{\Sigma(x_3x_4)}{N}$$

$$\text{(IV)}\quad \frac{\Sigma(x_1x_4)}{N} = a\frac{\Sigma(x_4)}{N} + b_{12\cdot34}\frac{\Sigma(x_2x_4)}{N} + b_{13\cdot24}\frac{\Sigma(x_3x_4)}{N} + b_{14\cdot23}\frac{\Sigma(x_4^2)}{N}$$

where x_1, x_2, x_3, x_4, represent deviations from the respective means, $\overline{X}_1$ $\overline{X}_2$, $\overline{X}_3$ and $\overline{X}_4$ but since the sum of the deviations about the arithmetic mean is zero

$$\frac{\Sigma(x_1)}{N} = 0,\ \frac{\Sigma(x_2)}{N} = 0,\ \frac{\Sigma(x_3)}{N} = 0,\ \frac{\Sigma(x_4)}{N} = 0$$

and[1]

$$\sigma_2 = \sqrt{\frac{\Sigma(x_2^2)}{N}},\ \sigma_3 = \sqrt{\frac{\Sigma(x_3^2)}{N}},\ \sigma_4 = \sqrt{\frac{\Sigma(x_4^2)}{N}}$$

or

$$\sigma_2^2 = \frac{\Sigma(x_2^2)}{N},\ \sigma_3^2 = \frac{\Sigma(x_3^2)}{N},\ \sigma_4^2 = \frac{\Sigma(x_4^2)}{N}$$

while

$$\frac{\Sigma(x_1x_2)}{N} = p_{12}\ \text{(the product moment)}^2$$

$$\frac{\Sigma(x_1x_3)}{N} = p_{13}\ \text{etc.}$$

where the value of the product moment may be computed from

$$p_{12} = \frac{\Sigma(X_1X_2)}{N} - \frac{\Sigma(X_1)}{N}\quad\frac{\Sigma(X_2)}{N}$$

The "normal" equations will now read

$$p_{12} = b_{12\cdot34}\,\sigma_2^2 + b_{13\cdot24}\,p_{23} + b_{14\cdot23}\,p_{24}$$

$$p_{13} = b_{12\cdot34}\,p_{23} + b_{13\cdot24}\,\sigma_3^2 + b_{14\cdot23}\,p_{34}$$

$$p_{14} = b_{12\cdot34}\,p_{24} + b_{13\cdot24}\,p_{34} + b_{14\cdot23}\,\sigma_4^2$$

[1]See page 34
[2]See page 81

TECHNICAL APPENDIX VIII

STANDARD ERROR OF ESTIMATE— MULTIPLE CORRELATION

The formula for the standard error for multiple correlation may be derived in the same fashion as the "least squares" formula for simple correlation

$$S^2_{1.23} = \frac{\Sigma(X_1^2) - b_{12.34}\Sigma(X_1X_2) - b_{13.24}\Sigma(X_1X_3) - b_{14.23}\Sigma(X_1X_4)}{N}$$

reducing this to deviations from the respective means this will read

$$S^2_{1.23} = \frac{\Sigma(x_1^2)}{N} - b_{12.34}\frac{\Sigma(x_1x_2)}{N} - b_{13.24}\frac{\Sigma(x_1x_3)}{N} - b_{14.23}\frac{\Sigma(x_1x_4)}{N}$$

or

$$S^2_{1.23} = \sigma^2_1 - b_{12.34}\,p_{12} - b_{13.24}\,p_{13} - b_{14.23}\,p_{14}$$

TECHNICAL APPENDIX IX

DERIVATION OF STANDARD ERROR OF[1] THE ARITHMETIC MEAN

N random samples of n items each are drawn and the individual values expressed as deviations from the true arithmetic mean of the universe. This may be written as follows:

Item Number	Sample #1	Sample #2	Sample #3	Sample N
1	x'	x'	x'	x'
2	x''	x''	x''	x''
3	x'''	x'''	x'''	x'''
.				
n	x_n	x_n	x_n	x_n
	Σx_1	Σx_2	Σx_3	Σx_4

If items #1 and #2 of each sample are added

$$x_{1+2} = x' + x''$$

But the standard deviation is equal to

$$\sigma = \sqrt{\frac{\Sigma(x^2)}{N}}$$

[1] This derivation follows Ezekial's in *Methods of Correlation Analysis*

or in this instance

$$\sigma = \sqrt{\frac{\Sigma(x^2_{1+2})}{N}}$$

or

$$\sigma_{1+2}{}^2 = \sqrt{\frac{\Sigma(x^2_{1+2})}{N}}$$

but

$$x_{1+2}{}^2 = (x' + x'')^2$$
$$= (x')^2 + 2(x'x'') + (x'')^2$$

and

$$\Sigma(x_{1+2}{}^2) = \Sigma(x'^2) + 2\Sigma(x'x'') + \Sigma(x''^2)$$

Since the successive items of the sample are drawn at random they are uncorrelated ($r = 0$) and therefore $\Sigma(x'x'') = 0$.

$$\therefore \Sigma(x^2_{1+2}) = \Sigma(x'^2) + \Sigma(x''^2)$$

or after dividing by N

$$\sigma^2_{1+2} = \sigma_{x'}{}^2 + \sigma_{x''}{}^2$$

But as the number (N) of the samples is increased $\sigma_{x'}$ will tend to approach the standard deviation of the universe from which the samples were drawn as will in a similar manner $\sigma_{x''}$

or $\sigma_{x'} = \sigma_x$

and $\sigma_{x''} = \sigma_x$ when N is very large

$$\therefore \sigma_{1+2}{}^2 = 2\sigma_x{}^2$$

For the sum of the first three items

$$\sigma^2_{1+2+3} = \sigma_{x'}{}^2 + \sigma_{x''}{}^2 = \sigma_{x'''}{}^2 = 3\sigma^2_x$$

when N is large

And for the sum of n items

$$\sigma_{1+2+3\ldots n}{}^2 = \sigma_{x'}{}^2 + \sigma_{x''}{}^2 + \ldots + \sigma_{x^n}{}^2$$
$$= n\sigma_x{}^2 \quad \text{when } N \text{ is large}$$

Dividing all items and totals by n gives

$$\left(\frac{x'}{N}\right)^2 = \frac{(x'^2)}{N}$$

and

$$\sigma^2_{\frac{x'}{n}} = \frac{\sigma_{x'}{}^2}{n^2}$$

and since $\sigma_{x'}$ tends to equal σ_x

$$\sigma_{\frac{x'}{n}}^2 = \frac{\sigma_x^2}{n^2} \qquad \text{when } N \text{ is large}$$

and for x' and x''

$$\sigma_{\frac{x'}{n}+\frac{x''}{n}}^2 = \frac{\sigma_x^2}{n^2} + \frac{\sigma_x^2}{n^2} = 2\frac{\sigma_x^2}{n^2} =$$

or for the sum of n items

$$\sigma_{\frac{x'}{n}+\frac{x''}{n}+\ldots+\frac{x^n}{n}}^2 = \frac{\sigma_x^2}{n^2} + \frac{\sigma_x^2}{n^2} + \ldots + \frac{\sigma_x^2}{n^2} = \frac{n\sigma_x^2}{n} = \frac{\sigma_x^2}{n}$$

but since $\Sigma\left(\frac{x}{n}\right)$ for each sample $= \bar{X}$

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{N}$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$$

where

σ_x is the standard deviation of the *universe* and not of the sample. However lacking this value the standard deviation of the sample (σ) is used as an estimate of this value.

INDEX